Joint COCO and Mapillary Workshop at ICCV 2019: LVIS Challenge Track

Technical Report: Equalization Loss for Large Vocabulary Instance Segmentation

Jingru Tan
Tongji University
tjr120@tongji.edu.cn

Changbao Wang
SenseTime Research
wangchangbao@sensetime.com

Quanquan Li
SenseTime Research
liquanquan@sensetime.com

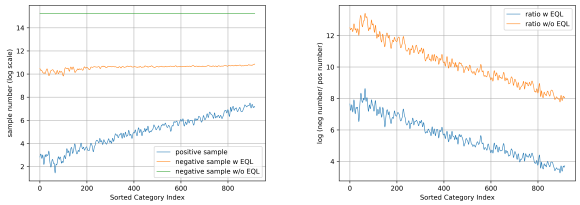
Junjie Yan
SenseTime Research
yanjunjie@sensetime.com

Abstract

Recent object detection and instance segmentation tasks mainly focus on datasets with relatively small set of categories, e.g. Pascal VOC with 20 classes and COCO with 80 classes. The new large vocabulary dataset LVIS brings new challenges to conventional methods. In this work we propose a equalization loss to solve the long tail of rare categories problem. Combined with exploiting the data from detection datasets to alleviate the effect of missing-annotation problem during the training, our method achieves 5.1% AP gain and 11.4% AP gain of rare categories on LVIS benchmark without any bells and whistles compared to Mask R-CNN baseline. Finally we achieve 28.9 mask AP on test set of the LVIS Challenge 2019.

1. Introduction

Different from preceding instance segmentation datasets such as COCO [7], the large vocabulary instance segmentation dataset LVIS [2] poses new challenges. First, unbalanced data distribution of categories leads to serious performance degradation of rare categories. Second, LVIS is not exhaustively annotated with all categories therefore unlabeled object instances will be treated as background and will generate incorrect supervision signal. Recent state-of-the-art methods show poor performance on LVIS [2], especially for the rare categories. In this work, we focus on these two problems. For the long-tail problem, we propose a new loss function to improve the performance of rare categories, which will be described in Section 2. For the missing-annotation problem, we provide simple but effective strategies to utilize object detection data and annotations, which will be described in Section 3.



(a) (b)

Figure 1: The effect of equalization loss on positive and negative samples. (a) shows the number of positive and negative samples for each category; (b) displays the ratio of the number of negative samples to the number of positive samples. Categories are sorted by their frequency.

2. Equalization Loss

Our work is based on the state-of-the-art instance segmentation framework Mask R-CNN [3] with two modifications. First, to alleviate the competition between categories, we replace softmax cross-entropy loss with sigmoid cross-entropy loss for the box classification. Second, to reduce the computation and memory cost, we use class-agnostic mask prediction for the mask head instead of class-specific mask prediction in origin paper [2].

During the box prediction stage of Mask R-CNN, for proposal R assigned with category c, the sigmoid cross-entropy loss of classification branch can be computed as:

$$L_{cls} = -\sum_{j=1}^{C} log(p_j^*), \tag{1}$$

which

$$p_j^* = \begin{cases} p_j & \text{if } j = c \\ 1 - p_j & \text{otherwise,} \end{cases} \quad (2)$$

where C is the total number of categories, p_j is the predicted confidence for category j. This loss function requires that for a given proposal, it should try to predict only one category. However, in LVIS, one object can be annotated with multiple categories, there is no strict boundary between some categories. Meanwhile, since the annotations of rare and common categories are much less than that of frequent categories, predictions for rare and common categories are suppressed for almost all the time using Equation 1 and 2. In another word, a positive sample of one category can be seen as a negative sample for other categories at the same time. Those negative signals have a marked impact to categories with scarce annotations, i.e. rare and common. We claim that less punishment to the rare the common objects helps alleviate the two problems mentioned above, so we introduce a novel equalization loss. It adds an additional weight $w \in \mathbb{R}^C$ to the origin sigmoid loss function. Given a proposal R, we compute w as follows: if the proposal R is negative, w is set to 1 for all index j; For a positive R, w is set to 0 if $j < \lambda$ and its category c is not in the union of positive category set and negative category set. The equalization loss is formulated as:

$$L_{EQL} = -\sum_{j=1}^{C} w_j log(p_j^*), \quad (3)$$

which

$$w_j = \begin{cases} 0 & \text{if } c > 0 \text{ and } f_j < \lambda \text{ and } j \notin S_P \cup S_N \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

where c is the category for proposal R, f_j is the frequency of category j, S_P and S_N are the positive and negative category sets of the ground truth annotations of the image. Since the categories are classified to frequent, common and rare in LVIS, in our experiments, we empirically set λ to ignore all rare and common categories. A more detailed parameter search of λ may bring further improvements.

The effect of the proposed loss is shown in Figure 1. As we can see, it alleviate the imbalance problem between positive and negative samples.

3. Exploiting Data of Object Detection

Since LVIS is not exhaustively annotated with all categories and data of rare categories are quite scarce, we utilize additional public datasets and provide several strategies.

3.1. COCO Ignore

If a proposal is assigned to negative sample because of missing-annotation, (miss-annotations problem comes from unknown categories and "not exhaustive" cases, not annotations errors), the model will get an incorrect updating signal, which will influence the training and degrades the performance. Since LVIS and COCO dataset share same set of images, we utilize the bounding box annotation from COCO dataset. During training, we calculate the overlaps between negative proposals and COCO ground truth bounding boxes. Figure 2 shows some examples of the missing-annotation situation. For those IoU larger than 0.5, we decrease the weight to β. Since this strategy changes the total loss scale, we double the loss weight of the box head and set β to 0.5 rather than 0.

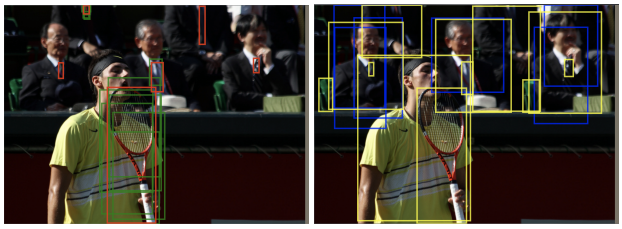


Figure 2: Examples of miss-annotation. Red and green boxes are LVIS annotated boxes and positive proposals, yellow and blue boxes are COCO annotated boxes and negative proposals. Thoses blue negative proposals will give incorrect updating signals to the model.

3.2. COCO Pre-training

Transfer learning is helpful for relatively small dataset, and LVIS share the same image sets with COCO, so it's intuitive to train a detector in LVIS using COCO-pretrained model instead of ImageNet-pretrained model. We pretrain our model on COCO with instance segmentation annotations and then fine-tune on LVIS.

3.3. OpenImage

OpenImage [5] is a large datasets with 600 object categories. LVIS shares 110 categories with OpenImage. We add corresponding (about 20k images) images to LVIS train set. Only bounding-box level annotations is used, so the losses of mask branch are ignored for those images of OpenImage dataset.

4. Experiments

We perform experiments on LVIS dataset, which contains 1230 categories in release v0.5. The evaluation metric is AP across IoU threshold from 0.5 to 0.95. We train our model on 60k `train` images and test it on 5k `val` set. We also reported our results on 20k `test` images.

Model	AP	AP_r	AP_c	AP_f
ResNet50-Softmax [2]	**21.0**	3.2	21.3	27.7
ResNet50-Softmax*	20.8	5.6	**21.5**	25.6
ResNet50-Sigmoid	20.1	**6.5**	19.9	25.4

Table 1: Baseline model results on LVIS val v0.5. ResNet50-Softmax* is our re-implementation. The metrics are mask AP and subscripts 'r', 'c', and 'f' stands for rare, common, and frequent category.

Model	AP	AP_r	AP_c	AP_f
Baseline Sigmoid Loss	20.1	6.5	19.9	25.4
Equalization Loss	22.8	10.1	25.0	25.1

Table 2: Comparison of Equalization Loss and naive sigmoid loss. All model are trained using ResNet-50 Mask R-CNN.

EQL	RS	PR	IG	AP	AP_r	AP_c	AP_f
				20.1	6.5	19.9	25.4
	✔			21.3	12.2	21.5	24.7
✔				22.8	10.1	25.0	25.1
✔	✔			23.3	15.3	24.8	24.5
✔		✔		23.9	11.7	26.0	26.1
✔			✔	23.5	14.8	25.0	25.2
✔	✔	✔		**25.2**	**17.9**	**26.8**	**26.2**

Table 3: Experiment results of **EQL**(Equalization Loss), **RS**(Resampling), **PR**(COCO Pretrain), **IG**(COCO Ignore)

4.1. Implementation Details

We implement standard Mask R-CNN equipped with FPN [6] as our model. Training images are resized such as its shorter edge is 800 pixels. No other augmentation is used except horizontal flipping. RPN samples 256 anchors with 1:1 ratio of positive to negative. RoIAlign [3] is adopted to extracted features of proposals. R-CNN head samples 512 proposals per image, with 1:3 ratio of positive to negative. Though class-specific mask prediction achieves better performance, we use class-agnostic regime due to memory and computation cost for mask branch. In testing, the score threshold is reduced from 0.05 to 0.0, and top 300 bounding boxes are remained as detection results. Other settings are kept the same as origin implementation if not mentioned.

4.2. Ablation Study

In this section, we perform ablation studies among Equalization Loss, COCO Ignore, COCO pre-training and class-aware resampling. We implement Mask R-CNN with ResNet-50 and replace the conventional softmax cross-entropy loss with sigmoid cross entropy loss in box head as our baseline. Comparisons of sigmoid loss and softmax loss used in origin paper [2] are shown in Table 1.

Equalization Loss The experiments results are shown in Table 2. Comparing with the sigmoid cross-entropy loss, our method can lead to a significant improvement from 20.1% to 22.8%, especially on rare and common categories.

COCO Ignore We study the effectiveness COCO Ignore. As shown in Table 3, COCO Ignore can significantly improve the AP_r by 4.7% (10.1% to 14.8%), and lead an 0.7% overall AP improvement.

COCO Pre-training We demonstrate the effectiveness of COCO Pre-training in Table 3. It brings consistent performance gain on three category groups.

Resampling We also implemented the class-aware data resampling method which is proposed in [2]. We find that our equalization loss is compatible with resampling, EQL can further increase AP by 2% with resampling (from 21.3% to 23.3%). Combining these two method can improve the overall AP by 3.2%.

5. LVIS Challenge 2019

We add several enhancements on our model for the challenge, results are shown in Table 4. With those enhancements, we achieve 36.4 and 28.9 Mask AP on `val` and `test` set respectively which is demonstrated in Table 5.

Challenge Baseline We replace ResNet50 with ResNeXt-101-64x-4d [9] and use synchronized batch normalization [8] in backbone and heads. Also deformable convolution is adopted [1] in stage3, stage4 and stage5 of the model. We use multi-scale training and images are resized to range from 400 to 1400, and the longer edge is set to 1400. All these enhancements achieve a AP of 30.1%, shown in Table 4.

Multi-scale Testing We apply multi-scale testing on both bounding box and segmentation results. The testing scales are set to 600, 800, 1000, 1200, 1400.

Expert Model We train two large model on COCO and OpenImage respectively, and do testing on LVIS test set. Shared categories detection results are used for ensemble. Expert models of COCO and OpenImage improve the AP by 0.2 and 0.2 respectively.

Re-scoring Ensemble Due to the class imbalance problem, the detection scores for rare and common categories are much lower than that of frequent categories. We observed that ensemble degrades recall of rare and common categories dramatically because of their low scores compared to frequent ones. To solve this, the predictions are sorted by their scores and category frequency jointly. For prediction d_r of rare categories, and d_f of frequent categories, d_r is prior to d_f if $score_r + \alpha > score_f$, so do the common vs the frequent. In our experiments, we set α to 0.1 for the rare and 0.05 for the common, respectively.

Model	AP	AP_r	AP_c	AP_f
Challenge Baseline	30.1	19.3	31.8	32.3
+SE154 [4]	30.8	19.7	32.2	33.4
+OpenImage Data	31.4	21.5	33.1	33.3
+Multi-scale box testing	32.3	20.5	34.7	34.2
+RS Ensemble + Expert Model	35.1	24.8	37.5	36.3
+Multi-scale mask testing	**36.4**	**25.5**	**38.6**	**38.1**

Table 4: Experiment results of different tricks. RS Ensemble stands for Rescoring Ensemble.

	eval.set	AP	AP_r	AP_c	AP_f
[2]	`val`	27.1	15.6	27.5	31.4
Ours	`val`	**36.4**	**25.5**	**38.6**	**38.1**
[2]	`test`	20.5	9.8	21.1	30.0
Ours	`test`	**28.9**	**17.7**	**30.8**	**36.7**

Table 5: Final results on val and test.

References

[1] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[2] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 1, 3, 4

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 3

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[5] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 2

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[8] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6181–6189, 2018. 3

[9] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3