南京航空航天大學

Nanjing University of Aeronautics and Astronautics

CVPR

Conference on Computer Vision and Pattern Recognition

Two Articles about Loss function for long-tail distribution

Equalization Loss for Long-Tailed Object Recognition

Jingru Tan[1] Changbao Wang[2] Buyu Li[3] Quanquan Li[2]
Wanli Ouyang[4] Changqing Yin[1] Junjie Yan[2]
[1]Tongji University [2]SenseTime Research [3]The Chinese University of Hong Kong
[4]The University of Sydney, SenseTime Computer Vision Research Group, Australia
{tjr120,yinchangqing}@tongji.edu.cn, {wangchangbao,liquanquan,yanjunjie}@sensetime.com
byli@ee.cuhk.edu.hk, wanli.ouyang@sydney.edu.au

CVPR 2020

Motivation

The problem of the long-tailed distribution of the categories is a great challenge to the learning of object detection models, especially for the rare categories. So the rare categories can be easily overwhelmed by the majority categories during training and are inclined to be predicted as negatives. Thus the conventional object detectors trained on such an extremely unbalanced dataset suffer a great decline.

Motivation

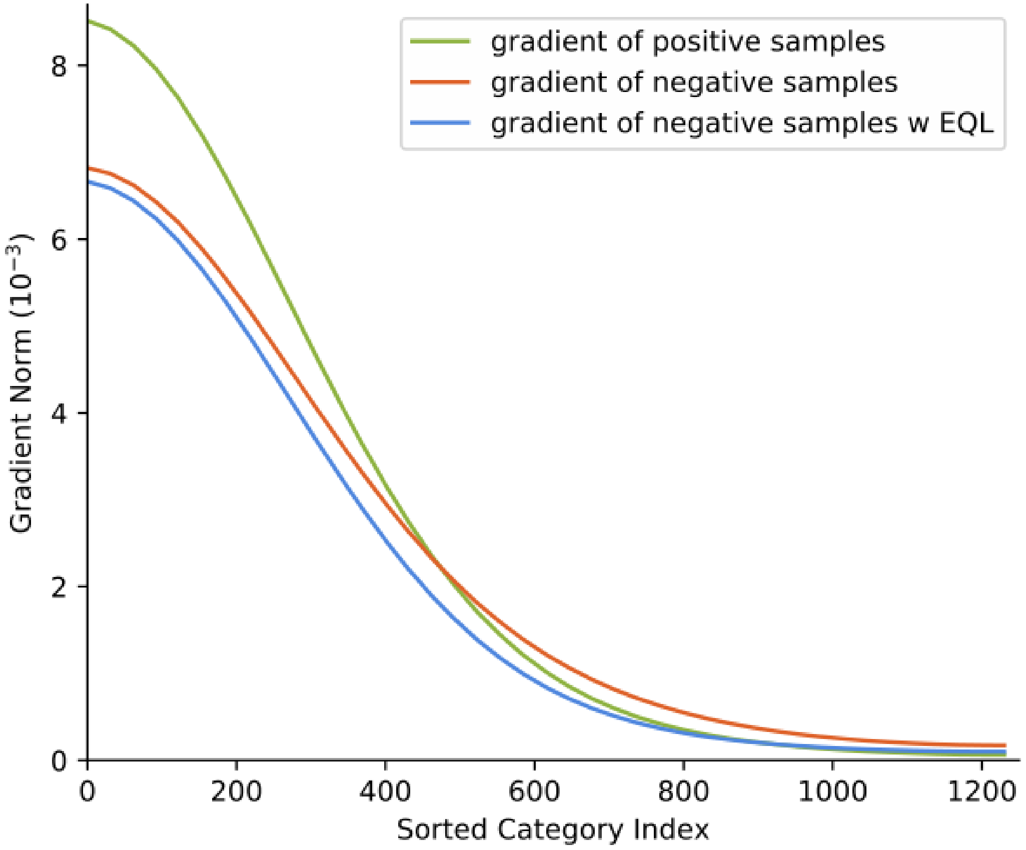


Figure 1: The overall gradient analysis on positive and negative samples. We collect the average L_2 norm of gradient of weights in the last classifier layer. Categories' indices are sorted by their instance counts. Note that for one category, proposals of all the other categories and the background are negative samples for it.

Each positive sample of one category can be seen as a negative sample for other categories, making the tail categories receive more discouraging gradients.

cross-entropy loss

Softmax cross-entropy loss $$L_{SCE} = -\sum_{j=1}^{C} y_j \log(p_j) \quad (1)$$ $$y_j = \begin{cases} 1 & \text{if } j = c \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

sigmoid cross-entropy loss $$L_{BCE} = -\sum_{j}^{C} \log(\hat{p_j}) \quad (3)$$ $$\hat{p_j} = \begin{cases} p_j & \text{if } y_j = 1 \\ 1 - p_j & \text{otherwise} \end{cases} \quad (4)$$

The derivative of the L_{BCE} and L_{SCE} with respect to network's output z in sigmoid cross entropy

$$\frac{\partial L_{cls}}{\partial z_j} = \begin{cases} p_j - 1 & \text{if } y_j = 1 \\ p_j & \text{otherwise} \end{cases} \quad (5)$$

equalization loss (EQL)

Equalization loss
$$L_{EQL} = -\sum_{j=1}^{C} w_j log(\hat{p_j}) \quad (6)$$

$$w_j = 1 - E(r)T_\lambda(f_j)(1 - y_j) \quad (7)$$

Tail Ratio
$$TR(\lambda) = \frac{\sum_j^C T_\lambda(f_j)N_j}{\sum_j^C N_j} \quad (8)$$

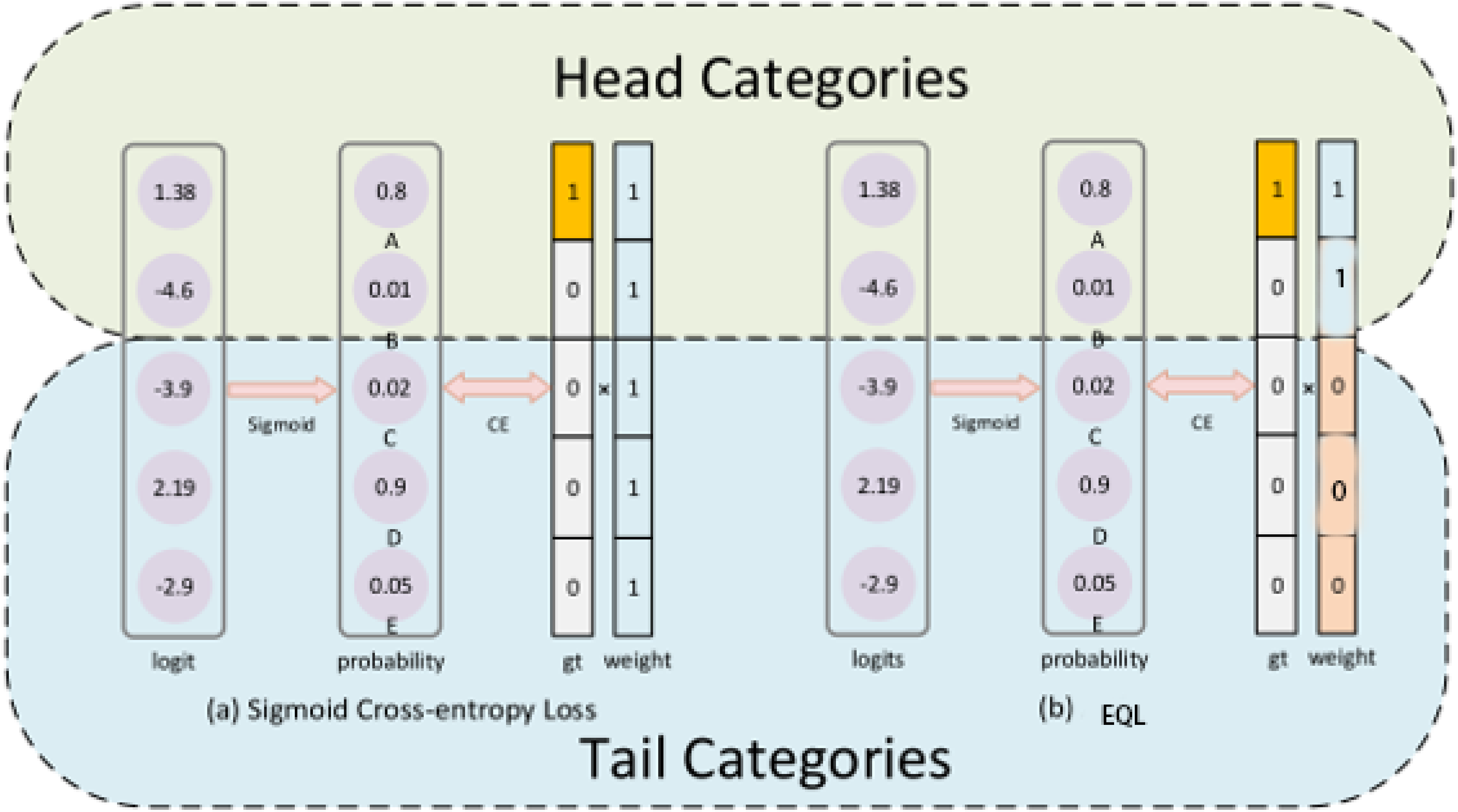
E(r) outputs 1 when r is a foreground region proposal and 0 when it belongs to background

In summary, there are two particular designs in equalization loss function:

1) We ignore the discouraging gradients of negative samples for rare categories whose quantity frequency is under a threshold.

2) We do not ignore the gradients of background samples. If all the negative samples for the rare categories are ignored, there will be no negative samples for them during training, and the learned model will predict a large number of false positives.

equalization loss (EQL)



Softmax equalization loss (SEQL)

Extend to Image Classification

Softmax equalization loss (SEQL)

$$L_{SEQL} = -\sum_{j=1}^{C} y_j \log(\tilde{p_j}) \tag{9}$$

$$\tilde{p_j} = \frac{e^{z_j}}{\sum_{k=1}^{C} \tilde{w_k} e^{z_k}} \tag{10}$$

$$\tilde{w_k} = 1 - \beta T_\lambda(f_k)(1 - y_k) \tag{11}$$

Experiments on Object Recognition

南京航空航天大学 Nanjing University of Aeronautics and Astronautics

1st place in the LVIS Challenge 2019/2020

	Backbone	EQL	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_{bbox}
Mask R-CNN	R-50-C4	✗	19.7	32.5	20.3	7.9	21.1	22.8	20.3
		✓	22.5	36.6	23.5	14.4	24.9	22.6	23.1
Mask R-CNN	R-101-C4	✗	21.8	35.6	22.7	10.5	23.4	24.2	22.9
		✓	24.1	38.7	25.6	15.8	26.8	24.1	25.6
Mask R-CNN	R-50-FPN	✗	20.1	32.7	21.2	7.2	19.9	25.4	20.5
		✓	22.8	36.0	24.4	11.3	24.7	25.1	23.3
Mask R-CNN	R-101-FPN	✗	22.2	35.3	23.4	9.8	22.6	26.5	22.7
		✓	24.8	38.4	26.8	14.6	26.7	26.4	25.2
Cascade Mask R-CNN	R-50-FPN	✗	21.1	33.3	22.2	6.3	21.6	26.5	21.1
		✓	23.1	35.7	24.3	10.4	24.5	26.3	23.1
Cascade Mask R-CNN	R-101-FPN	✗	21.9	34.3	23.2	6.0	22.3	27.7	24.7
		✓	24.9	37.9	26.7	10.3	27.3	27.8	27.9

Table 1: Results on different frameworks and models. All those models use class-agnostic mask prediction and are evaluated on LVIS v0.5 `val` set. AP is mask AP, and subscripts 'r', 'c' and 'f' stand for rare, common and frequent categories respectively. For equalization loss function, the λ is set as 1.76×10^{-3} to include all the rare and common categories.

Experiments on Object Recognition

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_S	AP_M	AP_L	AP_{bbox}
Sigmoid Loss	20.1	32.7	21.2	7.2	19.9	25.4	19.3	35.7	45.0	20.5
Softmax Loss	20.2	32.6	21.3	4.5	20.8	25.6	19.9	36.3	44.7	20.7
Class-aware Sampling [38]	18.5	31.1	18.9	7.3	19.3	21.9	17.3	32.1	40.9	18.4
Repeat Factor Sampling [15]	21.3	34.9	22.0	**12.2**	21.5	24.7	19.6	35.3	46.2	21.6
Class-balanced Loss [5]	20.9	33.8	22.2	8.2	21.2	25.7	19.8	36.1	46.4	21.0
Focal Loss [27]	21.0	34.2	22.1	9.3	21.0	**25.8**	19.8	36.5	45.5	21.9
EQL(Ours)	**22.8**	**36.0**	**24.4**	11.3	**24.7**	25.1	**20.5**	**38.7**	**49.2**	**23.3**

Table 5: Comparison with other methods on LVIS v0.5 `val` set. All experiments are performed based on ResNet-50 Mask R-CNN.

Ablation Studies on Object Recognition

Frequency Threshold λ:

$\lambda(10^{-3})$	TR(%)	AP	AP_r	AP_c	AP_f	AP_{bbox}
0	0	20.1	7.2	19.9	25.4	20.5
$0.176(\lambda_r)$	0.93	20.8	**11.7**	20.2	25.2	20.8
0.5	3.14	22.0	11.2	22.8	25.2	22.4
0.8	4.88	22.3	11.4	23.4	25.3	23.0
1.5	7.82	22.8	11.0	24.5	25.5	23.0
$1.76(\lambda_c)$	9.08	**22.8**	11.3	**24.7**	25.1	**23.3**
2.0	9.83	22.7	11.3	24.3	25.3	23.2
3.0	13.12	22.5	11.0	24.0	25.3	23.1
5.0	18.17	22.4	10.0	23.6	**25.7**	23.0

Table 2: Ablation study for different λ. λ_r is about 1.76×10^{-4}, which exactly includes all rare categories. λ_c is about 1.76×10^{-3}, which exactly includes all rare and common categories. When λ is 0, our equalization loss degenerates to sigmoid cross-entropy.

Ablation Studies on Object Recognition

Frequen Threshold Function $T_\lambda(f)$:

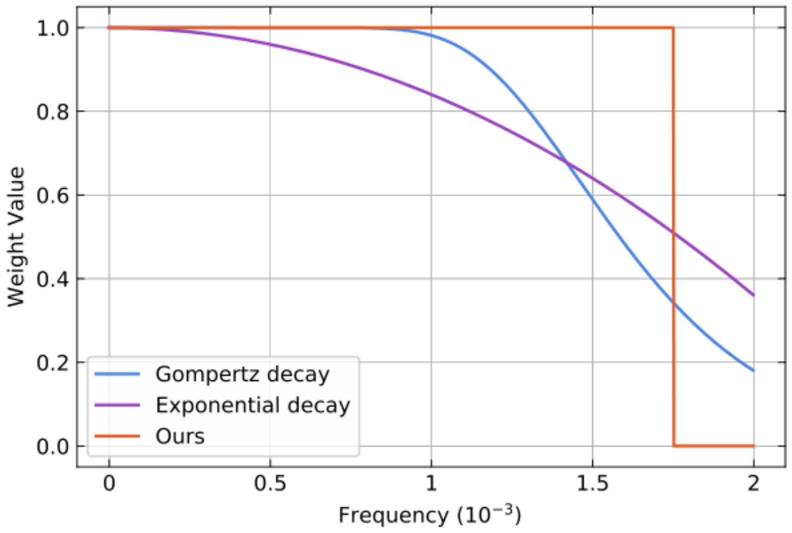


Figure 3: Illustration of different design for threshold function $T_\lambda(f)$.

	AP	AP_r	AP_c	AP_f	AP_{bbox}
Exponential decay	22.3	10.4	24.0	25.0	22.8
Gompertz decay	22.7	11.0	24.5	25.1	23.2
Ours	**22.8**	**11.3**	**24.7**	**25.1**	**23.3**

Exponential decay $y = 1 - (af)^n$ a= 400 and n= 2

Gompertz decay $y = 1 - ae^{-be^{-cf}}$ a= 1, b= 80, c= 3000

Ablation Studies on Object Recognition

E(r):

$E(r)$	AP	AP_r	AP_c	AP_f	AP_{bbox}
✗	22.2	**12.5**	24.7	23.1	22.7
✓	**22.8**	11.3	24.7	**25.1**	**23.3**

Table 4: Ablation study of Excluding Function $E(r)$. The top row is the results without using the term $E(r)$, and the bottom row is the results with it.

Experiments

Experiments on Open Images Detection

Method	AP	AP_1	AP_2	AP_3	AP_4	AP_5
SGM	48.13	59.86	51.24	49.31	46.51	33.72
CAS [38]	56.50	64.44	59.30	59.74	57.02	42.00
EQL(Ours)	**57.83**	**64.95**	**60.18**	**61.17**	**58.23**	**44.6**

Table 6: Results on OID19 `val` set based on ResNet-50. **SGM** and **CAS** stand for sigmoid cross-entropy and class-aware sampling. We sort all the categories by their image number and divide them into 5 groups. TR and λ is 3% and 3×10^{-4} respectively.

Experiments on Image Classification

Method	Acc@top1	Acc@Top5
Focal Loss† [27]	35.62	-
Class Balanced† [5]	36.23	-
Meta-Weight Net† [40]	37.91	-
SEQL(Ours)	**43.38**	**71.94**

Table 8: Results on CIFAR-100-LT `test` set based on ResNet-32 [18]. We use γ of 0.95 and λ of 3.0×10^{-3}. † means that the results are copied from origin paper [5, 40]. Imbalanced factor is 200.

Method	Acc@Top1	Acc@Top5
FSLwF† [11]	28.4	-
Focal Loss† [27]	30.5	-
Lifted Loss† [34]	30.8	-
Range Loss† [44]	30.7	-
OLTR† [30]	35.6	-
SEQL(Ours)	**36.44**	**61.19**

Table 10: Results on ImageNet-LT `test` set based on ResNet-10 [18]. The optimal γ and λ are 0.9 and 4.3×10^{-4} respectively. † means that the results are copied from origin paper [30]

Adaptive Class Suppression Loss for Long-Tail Object Detection

Tong Wang[1,2], Yousong Zhu[1,3], Chaoyang Zhao[1], Wei Zeng[4,5], Jinqiao Wang[1,2,6], Ming Tang[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing, China
[3] ObjectEye Inc., Beijing, China
[4] Peking University, Beijing, China
[5] Peng Cheng Laboratory, Shenzhen, China
[6] NEXWISE Co., Ltd., Guangzhou, China

{tong.wang,yousong.zhu,chaoyang.zhao,jqwang,tangm}@nlpr.ia.ac.cn
weizeng@pku.edu.cn

CVPR 2021

Motivation

Table 1: Experiments on LVIS with different groups.

Groups	mAP	AP_r	AP_c	AP_f
(0,5)[5,∞)	22.74	6.83	22.14	29.83
(0,50)[50,∞)	25.30	15.11	24.99	29.77
(0,500)[500,∞)	25.66	13.19	25.98	30.25
(0,5000)[5000,∞)	23.89	8.27	23.87	30.16

Motivation

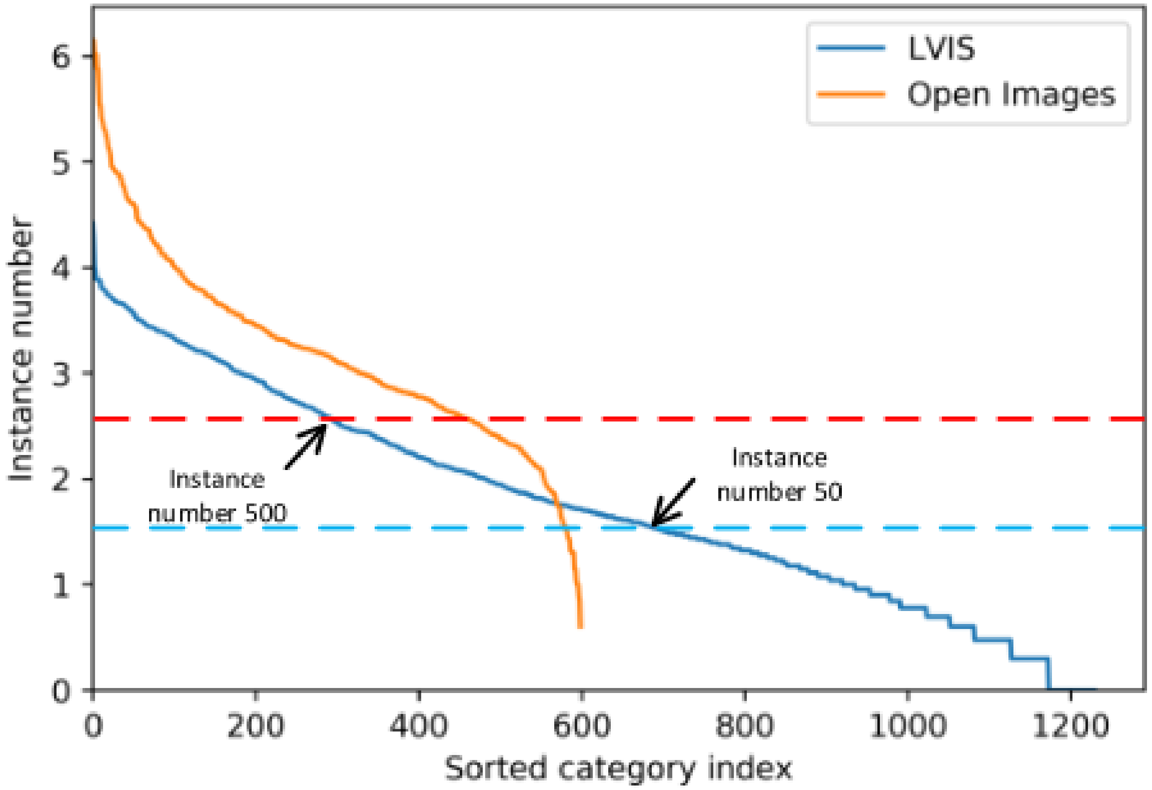


Figure 2: The data distribution of LVIS and Open Images dataset. The x-axis represents the sorted category index. Y-axis is the base-10 logarithm of the instance number.

Motivation

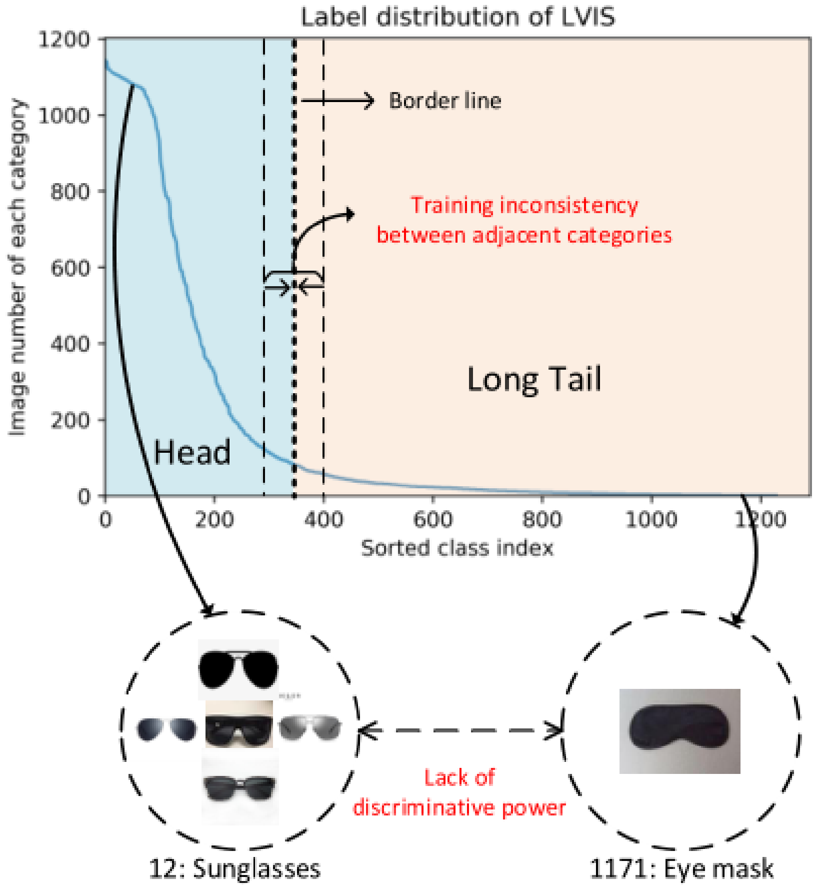


Figure 1: The label distribution of LVIS [11] dataset. The x-axis represents the sorted category index of LVIS. The y-axis is the image number of each category.

Adaptive Class Suppression loss (ACSL)

Adaptive Class Suppression Loss

$$L_{ACSL}(x_s) = -\sum_{i=1}^{C} w_i log(\hat{p_i})$$

$$w_i = \begin{cases} 1, & \text{if } i = k \\ 1, & \text{if } i \neq k \text{ and } p_i \geq \xi \\ 0, & \text{if } i \neq k \text{ and } p_i < \xi \end{cases}$$

$$\frac{\partial L_{ACSL}}{\partial z_i} = \begin{cases} p_i - 1, & \text{if } i = k \\ w_i p_i, & \text{if } i \neq k \end{cases}$$

Advantages:

ACSL takes the network learning status into consideration.

ACSL works in a more fine-grained sample level.

ACSL does not depend on the class distribution.

Adaptive Class Suppression loss (ACSL)

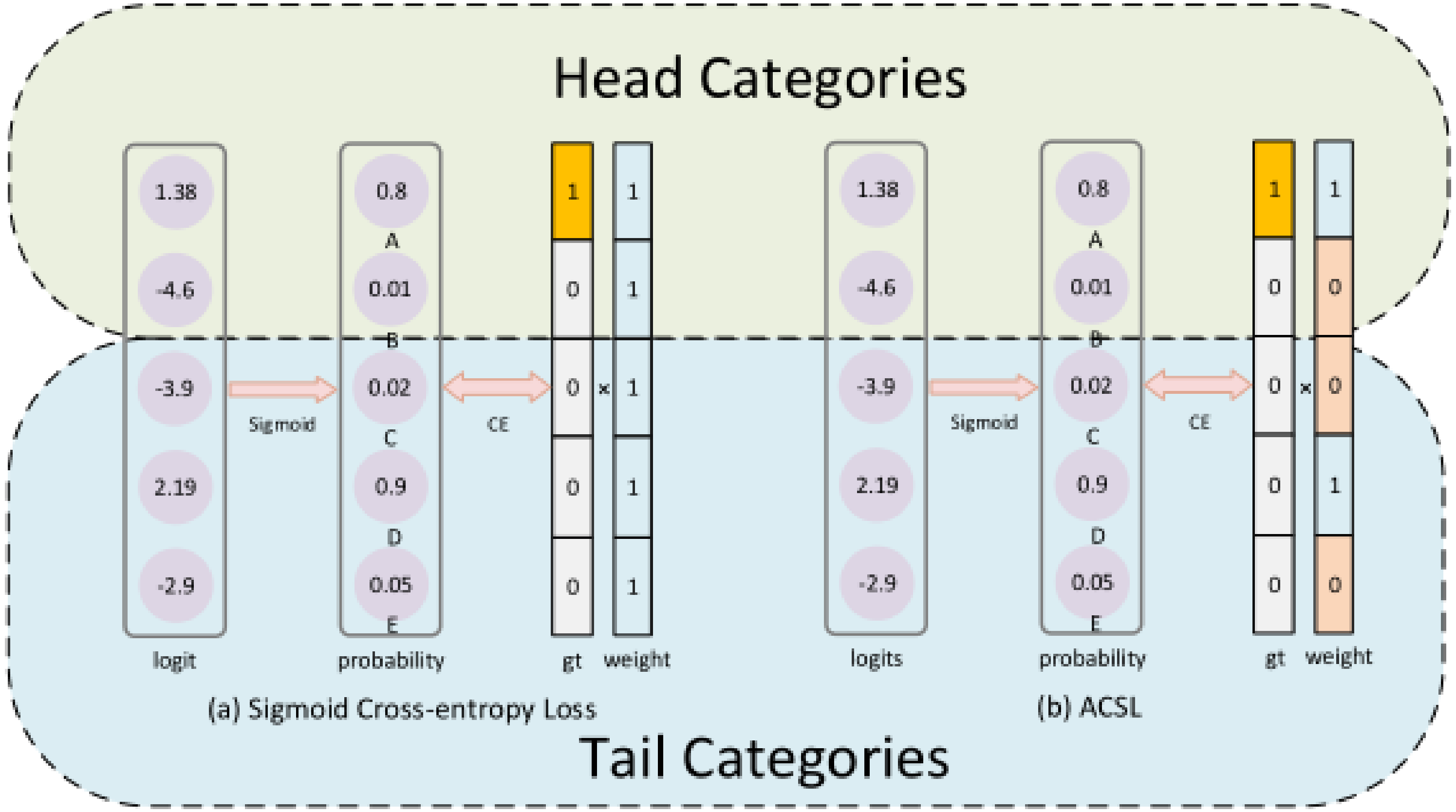
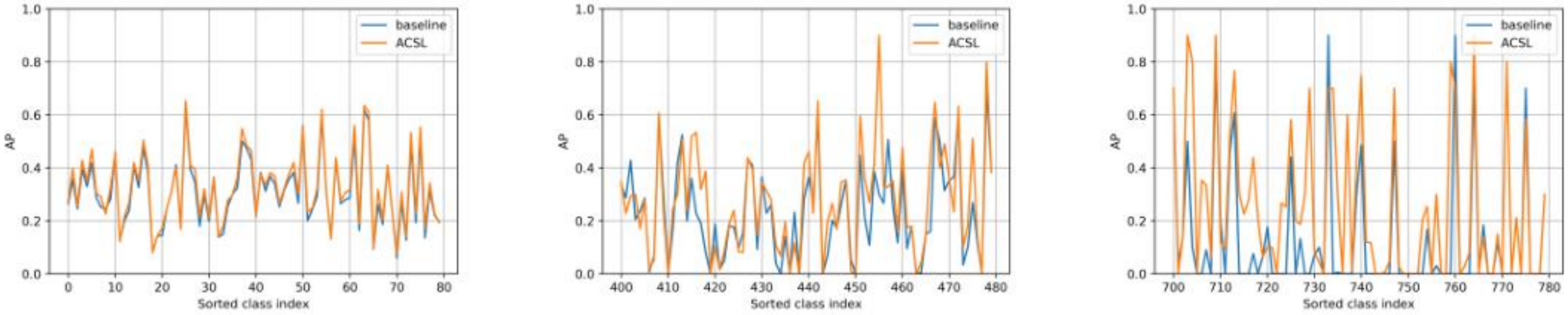


Figure 3: An illustration of Sigmoid Cross-entropy Loss and our proposed ACSL. The top two classes belong to head categories and the bottom three classes belong to tail categories. For ACSL, the hyper-parameter ξ is 0.7.

Experiments on LVIS



(a) The AP on frequent categories (b) The AP on common categories (c) The AP on rare categories

Figure 4: The AP of baseline and ACSL on frequent, common and rare categories, respectively. Both models are trained with ResNet50-FPN backbone. The x-axis is the sorted class index. The y-axis means the precision.

Experiments on LVIS

南京航空航天大学 Nanjing University of Aeronautics and Astronautics

Table 4: Comparison with state-of-the-art methods on LVIS-v0.5 *val* dataset. **Bold** numbers denote the best results among all models. "ms" means multi-scale testing.

Methods	backbone	mAP	AP_r	AP_c	AP_f	$AP@0.5$	$AP@0.75$	AP_s	AP_m	AP_l
Focal Loss [20]	ResNet-50	21.95	10.49	22.42	25.93	35.15	23.91	18.66	28.59	31.46
CBL [5]	ResNet-50	23.9	11.4	23.8	27.3	–	–	–	–	–
LDAM [2]	ResNet-50	24.5	14.6	25.3	26.3	–	–	–	–	–
RFS [11]	ResNet-50	24.9	14.4	24.5	29.5	41.6	25.8	19.8	30.6	37.2
LWS [15]	ResNet-50	24.1	14.4	24.4	26.8	–	–	–	–	–
SimCal [31]	ResNet-50	23.4	16.4	22.5	27.2	–	–	–	–	–
EQL [29]	ResNet-50	25.06	11.92	25.98	29.14	40.14	27.30	20.08	31.50	38.67
	ResNet-101	26.05	11.45	27.14	30.51	41.30	27.83	20.35	33.73	40.75
	ResNeXt-101-64x4d	28.04	15.03	29.14	31.87	44.06	30.07	22.19	34.52	42.97
BAGS [18]	ResNet-50	25.96	17.65	25.75	29.54	43.58	27.15	20.26	32.81	40.10
	ResNet-101	26.39	16.80	25.82	30.93	43.44	27.63	20.29	34.39	41.07
	ResNeXt-101-64x4d	27.83	18.78	27.32	32.07	45.83	28.99	21.92	35.65	43.11
ACSL (Ours)	ResNet-50	26.36	18.64	26.41	29.37	42.38	28.63	20.43	33.11	40.21
	ResNet-101	27.49	19.25	27.60	30.65	43.45	29.69	21.11	34.96	42.00
	ResNeXt-101-64x4d	28.93	**21.78**	28.98	31.72	45.54	31.19	22.16	35.81	43.43
	ResNet-50 (ms)	27.24	17.86	27.42	30.76	44.46	28.54	20.96	34.40	41.68
	ResNet-101 (ms)	28.23	17.42	28.40	32.32	44.73	30.13	21.86	35.43	44.06
	ResNeXt-101-64x4d (ms)	**29.47**	20.30	**29.45**	**33.15**	**46.82**	**31.55**	**22.52**	**37.32**	**45.51**

Moreover, the utilization of ACSL is not limited to a certain type of detector.

Experiments on Open Images

南京航空航天大学 Nanjing University of Aeronautics and Astronautics

Objects in Open Images have multiple labels, we train the models under multiple label setting.

Table 5: Experiments on Open Images with different backbones.

Backbone	Methods	*AP*
ResNet50-FPN	baseline	55.1
	ours	**60.3**
ResNet101-FPN	baseline	56.3
	ours	**61.6**
ResNet152-FPN	baseline	57.4
	ours	**62.8**

Table 6: The detailed precision on some of the tail categories of Open Images.

	Spa	Scr	Fac	Cas	Hor
img num	38	46	49	53	54
baseline	35.0	46.6	17.8	19.9	8.3
ACSL	**41.6(+6.6)**	**55.6(+9.0)**	**80.9(+63.1)**	**47.5(+27.6)**	**16.6(+8.3)**
	Slo	Obo	Squ	Bin	Ser
img num	103	93	97	109	106
baseline	25.0	22.2	29.1	42.7	40.2
ACSL	**45.0(+20)**	**83.3(+61.1)**	**50.3(+21.2)**	**61.5(+18.8)**	**73.2(+33)**

Table 7: Comparison with other methods on Open Images. All models are trained with ResNet50-FPN backbone and evaluated on 500 categories.

Method	AP
Class Aware Sampling [28]	56.50
Equalization Loss [29]	57.83
Ours	**61.70**

Ablation Study

Table 2: Experimental results of ACSL with different ξ.

	ξ	mAP	AP_r	AP_c	AP_f
baseline (1x)	–	21.18	4.30	20.09	29.28
baseline (2x)	–	22.28	7.38	22.34	28.17
ACSL	0.01	23.53	11.48	22.73	29.35
	0.1	25.11	16.04	24.72	29.22
	0.3	25.72	17.65	25.45	29.27
	0.5	26.08	18.61	25.85	29.36
	0.7	**26.36**	**18.64**	**26.41**	29.37
	0.9	25.99	17.25	26.0	**29.46**

Table 3: Results with larger backbones ResNet101, ResNeXt-101-64x4d and stronger detector Cascade R-CNN.

Models	Method	mAP	AP_r	AP_c	AP_f
Faster R101	baseline	22.36	3.14	21.82	**30.72**
	Ours	**27.49**	**19.25**	**27.60**	30.65
Faster X101	baseline	24.70	5.97	24.64	**32.26**
	Ours	**28.93**	**21.78**	**28.98**	31.72
Cascade R101	baseline	25.14	3.96	24.55	**34.35**
	Ours	**29.71**	**21.72**	**29.43**	33.26
Cascade X101	baseline	27.14	4.36	27.32	**36.03**
	Ours	**31.47**	**23.39**	**31.50**	34.66

Conclusion

1.
We propose a new statistic-free perspective to understand the long-tail distribution, thus significantly avoiding the dilemma of manual hard division.

2.
We present a novel adaptive class suppression loss (ACSL) that can effectively prevent the training inconsistency of adjacent categories and improve the discriminative power of rare categories.

3.
We conduct comprehensive experiments on long-tail object detection datasets L VIS and Open Images. ACSL achieves 5.18% and 5.2% improvements with ResNet50-FPN on L VIS and OpenImages respectively, which validates its effectiveness.

Other LOSS

$$L_{seesaw}(\mathbf{z}) = -\sum_{i=1}^{C} y_i \log(\widehat{\sigma}_i),$$

$$\text{with } \widehat{\sigma}_i = \frac{e^{z_i}}{\sum_{j\neq i}^{C} \mathcal{S}_{ij} e^{z_j} + e^{z_i}}.$$

The $\boldsymbol{S_{ij}}$ is obtained by multiplying the $\boldsymbol{M_{ij}}$ of the mitigation factor and the compensation factor $\boldsymbol{C_{ij}}$.

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

CVPR

Conference on Computer Vision and Pattern Recognition

Thanks for Listening