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

Abstract

Object recognition techniques using convolutional neural networks (CNN) have achieved great success. However, state-of-the-art object detection methods still perform poorly on large vocabulary and long-tailed datasets, e.g. *LVIS. In this work, we analyze this problem from a novel perspective: each positive sample of one category can be seen as a negative sample for other categories, making the tail categories receive more discouraging gradients. Based on it, we propose a simple but effective loss, named equalization loss, to tackle the problem of long-tailed rare categories by simply ignoring those gradients for rare categories. The equalization loss protects the learning of rare categories from being at a disadvantage during the network parameter updating. Thus the model is capable of learning better discriminative features for objects of rare classes. Without any bells and whistles, our method achieves AP gains of 4.1% and 4.8% for the rare and common categories on the challenging LVIS benchmark, compared to the Mask R-CNN baseline. With the utilization of the effective equalization loss, we finally won the 1st place in the LVIS Challenge 2019. Code has been made available at:* `https://github.com/tztztztztz/eql.detectron2`

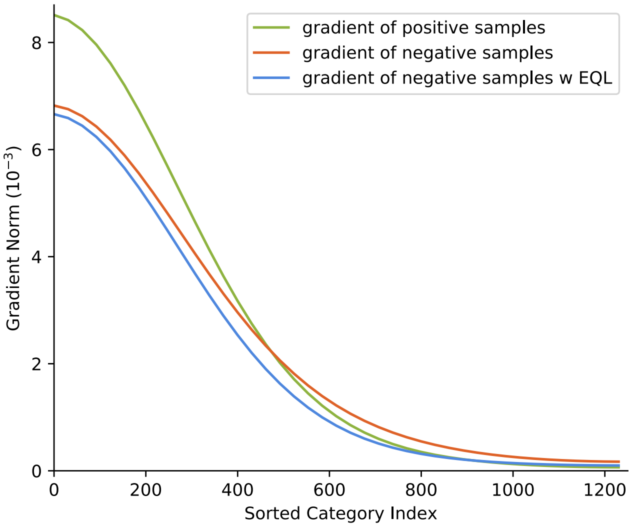

Figure 1: The overall gradient analysis on positive and negative samples. We collect the average L_2 norm of gradient of weights in the last classifier layer. Categories' indices are sorted by their instance counts. Note that for one category, proposals of all the other categories and the background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

1. Introduction

Recently, the computer vision community has witnessed the great success of object recognition because of the emerge of deep learning and convolutional neural networks (CNNs). Object recognition, which is a fundamental task in computer vision, plays a central role in many related tasks, such as re-identification, human pose estimation and object tracking.

Today, most datasets for general object recognition, *e.g.* Pascal VOC [10] and COCO [28], mainly collect frequently seen categories, with a large number of annotations for each class. However, when it comes to more practical scenarios, a large vocabulary dataset with a long-tailed distribution of category frequency (*e.g.* LVIS [15]) is inevitable. 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 (categories with very few samples). Note that for one category, all the samples of other categories including the background are regarded as negative samples. So the rare categories can be easily overwhelmed by the majority categories (categories with a large number of samples) during training and are inclined to be predicted as negatives. Thus the conventional object detectors trained on such an extremely unbalanced dataset

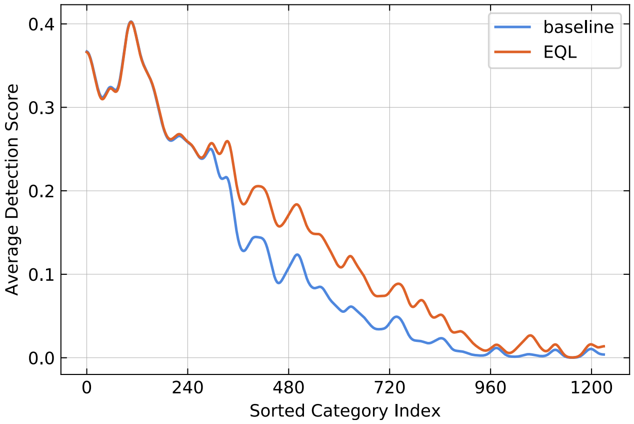

Figure 2: The predicted probabilities with EQL. The x-axis is the category index ordered by the instance number of each category, and the y-axis is the average predicted probability for positive proposals of each category.

suffer a great decline.

Most of the previous works consider the influence of the long-tailed category distribution problem as an imbalance of batch sampling during training, and they handle the problem mainly by designing specialized sampling strategies [2, 16, 32, 38]. Other works introduce specialized loss formulations to cope with the problem of positive-negative sample imbalance [27, 25]. But they focus on the imbalance between foreground and background samples so that the severe imbalance among different foreground categories remains a challenging problem.

In this work, we focus on the problem of extremely imbalanced frequencies among different foreground categories and propose a novel perspective to analyze the effect of it. As illustrated in Figure 1, the green and orange curves represent the average norms of gradients contributed by positive and negative samples respectively. We can see that for the frequent categories, the positive gradient has a larger impact than the negative gradient on average, but for the rare categories, the status is just the opposite. To put it further, the commonly used loss functions in classification tasks, *e.g*. softmax cross-entropy and sigmoid cross-entropy, have a suppression effect on the classes that are not the ground-truth one. When a sample of a certain class is utilized for training, the parameters of the prediction of the other classes will receive discouraging gradients which lead them to predict low probabilities. Since the objects of the rare categories hardly occur, the predictors for these classes are overwhelmed by the discouraging gradients during network parameters updating.

To address this problem, we propose a novel loss function, *equalization loss (EQL)*. In general, we introduce a weight term for each class of each sample, which mainly reduces the influence of negative samples for the rare categories. The complete formulation of equalization loss is presented in Section 3. With the equalization loss, the average gradient norm of negative samples decrease as shown in Figure 1 (the blue curve). And a simple visualization of the effect of EQL is shown in Figure 2, which illustrates the average predicted probabilities for the positive proposals of each category with (the red curve) and without (the blue curve) equalization loss. It can be seen that EQL significantly improves the performance on rare categories without harming the accuracy of frequent categories. With the proposed EQL, categories of different frequencies are brought to a more equal status during network parameter updating, and the trained model is able to distinguish objects of the rare categories more accurately.

Extensive experiments on several unbalanced datasets, *e.g*. Open Images [23] and LVIS [15], demonstrate the effectiveness of our method. We also verify our method on other tasks, like image classification.

Our key contributions can be summarized as follows: (1) We propose a novel perspective to analyze the long-tailed problem: the suppression on rare categories during learning caused by the inter-class competition, which explains the poor performance of rare categories on long-tailed datasets. Based on this perspective, a novel loss function, equalization loss is proposed, which alleviates the effect of the overwhelmed discouraging gradients during learning by introducing an ignoring strategy. (2) We present extensive experiments over different datasets and tasks, like object detection, instance segmentation and image classification. All experiments demonstrate the strength of our method, which brings a large performance boosting over common classification loss functions. Equipped with our equalization loss, we achieved the 1st place in the LVIS Challenge 2019.

2. Related Works

We first revisit common objection detection and instance segmentation. Then we introduce re-sampling, cost-sensitive re-weighting, and feature manipulation methods that are widely used to alleviate the class-unbalanced problem in long-tailed datasets.

Object Detection and Instance Segmentation. There are two mainstream frameworks for objection detection: single-stage detector [29, 36, 27] and two-stage detector [13, 12, 37, 26, 31]. While single-stage detectors achieve higher speed, most of state-of-the-art detectors follow the two-stage regime for better performance. The popular Mask R-CNN [17], which extends a mask head in the typical two-stage detector, provided promising results on many instance segmentation benchmarks. Mask Scoring R-CNN [21] introduced an extra mask score head to align the mask's score and quality. And Cascade Mask R-CNN [1] and HTC [3] further improved the performance by predicting the mask in a cascade manner.

Re-sampling Methods. One of the commonly used methods in re-sampling is oversampling [2, 16, 32], which randomly samples more training data from the minority classes, to tackle the unbalanced class distribution. Class-aware sampling [38], also called class-balanced sampling, is a typical technique of oversampling, which first samples a category and then an image uniformly that contains the sampled category. While oversampling methods achieve significant improvement for under-represented classes, they come with a high potential risk of overfitting. On the opposite of oversampling, the main idea of under-sampling [9] is to remove some available data from frequent classes to make the data distribution more balanced. However, the under-sampling is infeasible in extreme long-tailed datasets, since the imbalance ratio between the head class and tail class are extremely large. Recently, [22] proposed a decoupling training schema, which first learns the representations and classifier jointly, then obtains a balanced classifier by re-training the classifier with class-balanced sampling. Our method helps the model learn better representations for tail classes, so it could be complementary to the decoupling training schema.

Re-weighting Methods. The basic idea of re-weighting methods is to assign weights for different training samples. In an unbalanced dataset, an intuitive strategy is to weight samples based on the inverse of class frequency [41, 20] or use a smoothed version, inverse square root of class frequency [33]. Besides methods mentioned above which adjust the weight on class level, there are other studies focus on re-weighting on sample level. [27, 25] make the neural network to be cost-sensitive by increasing the weight for hard samples and decreasing the weight for easy samples, which can be seen as online versions of hard example mining technique [39]. Recently, Meta-Weight-Net [40] learns an explicit mapping for sample re-weighting. Different from the works above, we focus on the imbalance problem among different foreground categories. We propose a new perspective that the large number of negative gradients from frequent categories severely suppress the learning of rare categories during training. And we propose a new loss function to tackle this problem, which is applied to the sample level and class level simultaneously.

Feature Manipulation. There are also some works operating on the feature representations directly. Range Loss [44] enlarges inter-classes distance and reduces intra-classes variations simultaneously. [43] augments the feature space of tail classes by transferring the feature variance of regular classes that have sufficient training samples. [30] transfers the semantic feature representation from head to tail categories by adopting a memory module. However, designing those modules or methods is not a trivial task and makes the model harder to train. In contrast, our method is simpler and does not access the representation directly.

3. Equalization Loss

The central goal of our equalization loss is to alleviate the category quantity distribution imbalance problem for each category in a long-tailed class distribution. We start by revisiting conventional loss functions for classification, namely softmax cross-entropy and sigmoid cross-entropy.

3.1. Review of Cross-Entropy Loss

Softmax Cross-Entropy derives a multinomial distribution $\boldsymbol{p}$ over each category from the network outputs $\boldsymbol{z}$, and then computes the cross-entropy between the estimated distribution $\boldsymbol{p}$ and ground-truth distribution $\boldsymbol{y}$. The softmax cross-entropy loss L_{SCE} can be formulated as:

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

and C is the number of categories. Here, $\boldsymbol{p}$ is calculated by $Softmax(\boldsymbol{z})$. Note that the C categories include an extra class for background. In practice, $\boldsymbol{y}$ uses one-hot representation, and we have $\sum_{j=1}^{C} y_j = 1$. Formally, for the ground truth category c of a sample,

$$y_j = \begin{cases} 1 & \text{if } j = c \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Sigmoid Cross-Entropy estimates the probability of each category independently using C sigmoid loss functions. The ground truth label y_j only represents a binary distribution for category j. Usually, an extra category for background is not included. Instead, $y_j = 0$ will be set for all the categories when a proposal belongs to the background. So the sigmoid cross-entropy loss can be formulated as:

$$L_{BCE} = -\sum_{j}^{C} \log(\hat{p_j}) \tag{3}$$

where

$$\hat{p_j} = \begin{cases} p_j & \text{if } y_j = 1 \\ 1 - p_j & \text{otherwise} \end{cases} \tag{4}$$

Where p_j is calculated by $\sigma(z_j)$. The derivative of the L_{BCE} and L_{SCE} with respect to network's output $\boldsymbol{z}$ in sigmoid cross entropy shares the same formulation:

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

In softmax cross-entropy and sigmoid cross-entropy, we notice that for a foreground sample of category c, it can be regarded as a negative sample for any other category j. So the category j will receive a discouraging gradient p_j for model updating, which will lead the network to predict low probability for category j. If j is a rare category, the discouraging gradients will occur much more frequently than encouraging gradients during the iterations of optimization. The accumulated gradients will have a non-negligible impact on that category. Finally, even positive samples for category j might get a relatively low probability from the network.

3.2. Equalization Loss Formulation

When the quantity distribution of categories is fairly imbalanced, *e.g.* in a long-tailed dataset, the discouraging gradients from frequent categories have a remarkable impact on categories with scarce annotations. With commonly used cross-entropy losses, the learning of rare categories are easily suppressed. To solve this problem, we propose the equalization loss, which ignores the gradient from samples of frequent categories for the rare categories. This loss function aims to make the network training more fair for each class, and we refer it as equalization loss.

Formally, we introduce a weight term w to the original sigmoid cross-entropy loss function, and the equalization loss can be formulated as:

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

For a region proposal r, we set w with the following regulations:

$$w_j = 1 - E(r)T_\lambda(f_j)(1 - y_j) \tag{7}$$

In this equation, $E(r)$ outputs 1 when r is a foreground region proposal and 0 when it belongs to background. And f_j is the frequency of category j in the dataset, which is computed by the image number of the class j over the image number of the entire dataset. And $T_\lambda(x)$ is a threshold function which outputs 1 when $x < \lambda$ and 0 otherwise. λ is utilized to distinguish tail categories from all other categories and Tail Ratio (TR) is used as the criterion to set the value of it. Formally, we define TR by the following formula:

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

where N_j is the image number of category j. The settings of hyper-parameters of each part in Equation 7 are studied in Section 4.4.

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.

3.3. Extend to Image Classification

Since softmax loss function is widely adopted in image classification, we also design a form of Softmax Equalization Loss following our main idea. Softmax equalization loss (SEQL) can be formulated as:

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

where

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

and the weight term w_k is computed by:

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

where β is a random variable with a probability of γ to be 1 and $1 - \gamma$ to be 0.

Note that image classification is different from classification in object detection: each image belongs to a specific category, so there is no background category. Therefore, the weight term $\tilde{w_k}$ does not have the part $E(r)$ as in Equation 7. Therefore, we introduce β to randomly maintain the gradient of negative samples. And the influence of γ is studied in Section 6.

4. Experiments on LVIS

We conduct extensive experiments for equalization loss. In this section, we first present the implementation details and the main results on the LVIS dataset [15] in Section 4.2 and Section 4.3. Then we perform ablation studies to analyze different components of equalization loss in Section 4.4. In Section 4.5, we compare equalization loss with other methods. Details of LVIS Challenge 2019 will be introduced in Section 4.6.

4.1. LVIS Dataset

LVIS is a large vocabulary dataset for instance segmentation, which contains 1230 categories in current version v0.5. In LVIS, categories are divided into three groups according to the number of images that contains those categories: rare (1-10 images), common (11-100), and frequent (>100). We train our model on 57k `train` images and evaluate it on 5k `val` set. We also report our results on 20k `test` images. The evaluation metric is AP across IoU threshold from 0.5 to 0.95 over all categories. Different from COCO evaluation process, since LVIS is a sparse annotated dataset, detection results of categories that are not listed in the image level labels will not be evaluated.

	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.

4.2. Implementation Details

We implement standard Mask R-CNN [17] equipped with FPN [26] as our baseline model. Training images are resized such that its shorter edge is 800 pixels while the longer edge is no more than 1333. No other augmentation is used except horizontal flipping. In the first stage, RPN samples 256 anchors with a 1:1 ratio between the foreground and background, and then 512 proposals are sampled per image with 1:3 foreground-background ratio for the second stage. We use 16 GPUs with a total batch size 32 for training. Our model is optimized by stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0001 for 25 epochs, with an initial learning rate 0.04, which is decayed to 0.004 and 0.0004 at 16 epoch and 22 epoch respectively. Though class-specific mask prediction achieves better performance, we adopt a class-agnostic regime in our method due to the huge memory and computation cost for the large scale categories. Following [15], the threshold of prediction score is reduced from 0.05 to 0.0, and we keep the top 300 bounding boxes as prediction results. We make a small modification when EQL is applied on LVIS. Since for each image LVIS provide additional image-level annotations of which categories are in that image (positive category set) and which categories are not in it (negative category set), categories in EQL will not be ignored if they are in the positive category set or negative category set of that image, *i.e.* the weight term of Equation 7 will be 1 for those categories, even if they are rare ones.

4.3. Effectiveness of Equalization Loss

Table 1 demonstrates the effectiveness of equalization loss function over different backbones and frameworks. Besides Mask R-CNN, we also apply equalization loss on Cascade Mask R-CNN [1]. Our method achieves consistent improvement on all those models. As we can see from the table, the improvement mainly comes from the rare and common categories, indicating the effectiveness of our method on categories of the long-tailed distribution.

$\lambda(10^{-3})$	TR(%)	AP	AP$_r$	AP$_c$	AP$_f$	AP$_{bbox}$
0	0	20.1	7.2	19.9	25.4	20.5
0.176(λ_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(λ_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.

4.4. Ablation Studies

To better analyze equalization loss, we conduct several ablation studies. For all experiments we use ResNet-50 Mask R-CNN.

Frequency Threshold λ: The influence of different λ is shown in Table 2. We perform experiments of changing λ from 1.76×10^{-4}, which exactly split rare categories from all categories, to a broad range. We empirically find the proper λ locating in the space when $TR(\lambda)$ ranges from 2% to 10%. Results in Table 2 shows that significant improvement of overall AP as λ increases to include more tail

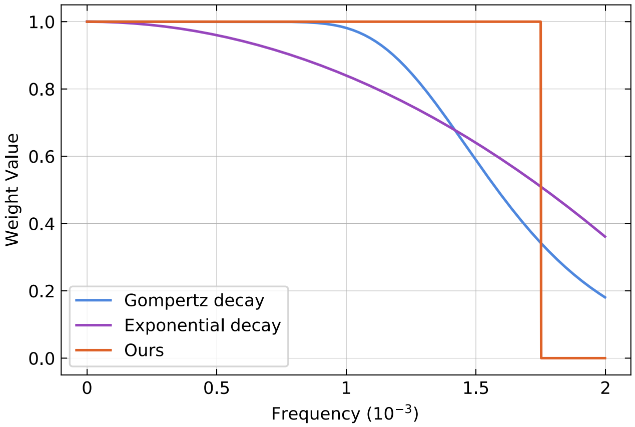

Figure 3: Illustration of different design for threshold function $T_\lambda(f)$.

	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**

Table 3: Ablation study for threshold function $T_\lambda(f)$. For a fair comparison, we compare the performance with their best hyper-parameters in multiple experiments.

$E(r)$	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.

categories. Meanwhile, the performance tends to degenerate when λ increases to include frequent categories. One advantage of equalization loss is that it has negligible effect on categories whose frequency is larger than a given λ. When $\lambda = \lambda_r$, AP_r improves significantly with marginal influence to AP_c and AP_f. And when $\lambda = \lambda_c$, AP_r and AP_c improve a lot while AP_f only degenerates slightly. We set λ to λ_c in all our experiments.

Threshold Function $T_\lambda(f)$: In Equation 7, we use $T_\lambda(f_j)$ to compute the weight of category j for a given proposal. Except for the proposed threshold function, $T_\lambda(f)$ can have other forms to calculate the weight for the categories with frequency under the threshold. As illustrated in Figure 3, we present and compare with another two designs: (1) Exponential decay function $y = 1 - (af)^n$, which computes the weight according to the power of category frequency. (2) Gompertz decay function $y = 1 - ae^{-be^{-cf}}$, which decays smoothly at the beginning and then decreases more steeply. We run multiple experiments for Exponential decay function and Gompertz decay function with different hyper-parameters and report the best results. The best hyper-parameter settings for Exponential decay function is $a = 400$ and $n = 2$ and for Gompertz decay function $a = 1, b = 80, c = 3000$. Table 3 shows that all of the three designs achieve fairly similar results, while both exponential decay and Gompertz decay function introduce more hyper-parameters to fit the design. Therefore, we use the threshold function in our method for its simpler format with less hyper-parameters and better performance.

Excluding Function $E(r)$: Table 4 shows the experiment results for EQL with or without $E(r)$. EQL without $E(r)$ means removing $E(r)$ from Equation 7, which will treat the foreground and background the same way. EQL with $E(r)$ means equalization loss only affects foreground proposals, as defined in Equation 7. Experiment results demonstrate the importance of $E(r)$. As we can see from the table, with $E(r)$, EQL achieves 0.6 points AP gain compared with EQL without $E(r)$. If $E(r)$ is discarded, although AP_r has an increase, AP_f drops dramatically, which causes the overall AP decline.

It is worth to notice that if we don't use $E(r)$, a large number of background proposals will be also ignored for rare and common categories, and the insufficient supervision from background proposals will cause extensive false positives. We visualize the detection results of an example image, which is shown in Figure 4. Without $E(r)$, more false positives are introduced, which are shown in red color. Both analysis and illustration above indicate that AP_r should decrease without $E(r)$, which is contradictory with the experiment results in Table 4. The reason is that according to LVIS evaluation protocol, if it is not sure whether category j is in or not in image I, all the false positives of category j will be ignored in image I. If category j is rare, the increased false positives are mostly ignored, which alleviates their influence. But the simultaneously increased true positives bring a direct increase in AP_r.

4.5. Comparison with Other Methods

Table 5 presents the comparison with other methods that are widely adopted to tackle the class imbalance problem. According to the table, re-sampling methods improve AP_r and AP_c at the sacrifice of AP_f, while re-weighting methods bring consistent gains on all categories but the overall improvement is trivial. The equalization loss improves AP_r and AP_c significantly with slight effect on AP_f, surpassing all other approaches.

4.6. LVIS Challenge 2019

With the help of the equalization loss, we finally won the 1st place on LVIS challenge held on COCO and Mapillary

	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.

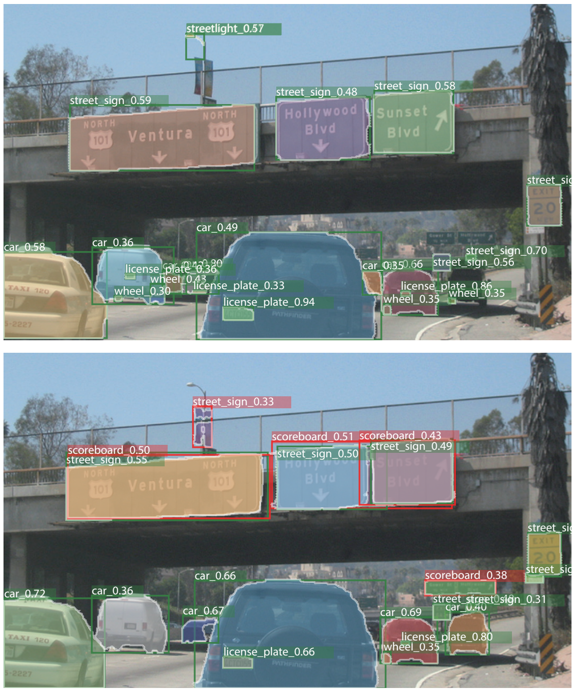

Figure 4: Illustration of the effect of excluding function $E(r)$. The upper and lower images correspond to using and removing $E(r)$ respectively. The true positives are drawn in green while false positives in red. We visualize the results with scores higher than 0.3 for better visualization.

Joint Recognition Challenge 2019. Combined with other enhancements, like larger backbone [19, 42], deformable convolution [6], synchronized batch normalization [35], and extra data, our method achieves a 28.9 mask AP on LVIS v0.5 `test` set, outperforming the ResNeXt-101 Mask R-CNN baseline (20.1%) by 8.4%. More details about our solution of challenge are described in Appendix A.

5. Experiments on Open Images Detection

Open Image dataset v5 is a large dataset of 9M images annotated with image-level labels and bounding boxes. In our experiments, we use the split of data and the subset of the categories of the competition 2019 for object detection track (OID19). The `train` set of OID19 contains 12.2M bounding boxes over 500 categories on 1.7M images, and the `val` contains about 10k images.

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.

According to Table 6, our method achieves a great improvement compared with standard sigmoid cross-entropy, outperforming class-aware sampling method by a significant margin. To better understand the improvement of our methods, we group all the categories by their image number and report the performance of each group. We can see that our method has larger improvements on categories with fewer samples. Significant AP gains on the group of fewest 100 categories are achieved compared with sigmoid cross-entropy and class-aware sampling (2.6 and 10.88 points respectively).

6. Experiments on Image Classification

To demonstrate the generalization ability of the equalization loss when transferring to other tasks. We also evaluate our method on two long-tailed image classification datasets, CIFAR-100-LT and ImageNet-LT.

Datasets. We follow exactly the same setting with [5] to generate the CIFAR-100-LT with imbalance factor of 200 [1]. CIFAR-100-LT contains 9502 images in `train` set, with 500 images for the most frequent category and 2 images

[1] `https://github.com/richardaecn/class-balanced-loss`

	$\lambda = 3.0 \times 10^{-3}$		$\lambda = 5.0 \times 10^{-3}$	
γ	Acc@top1	Acc@top5	Acc@top1	Acc@top5
0	41.33	67.75	41.33	67.75
0.75	42.08	70.03	42.26	69.95
0.9	43.12	71.50	**43.74**	**71.42**
0.95	**43.38**	**71.94**	43.30	72.31
0.99	42.44	71.44	42.49	72.07

Table 7: Varying γ and λ for SEQL. The accuracy is reported on CIFAR-100-LT `test` set. $\gamma = 0$ means we use softmax loss function.

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.

γ	Many	Medium	Few	Acc@top1	Acc@top5
0	**53.2**	27.5	8.0	34.7	58.7
0.5	52.5	28.7	9.8	35.2	59.7
0.75	52.1	30.7	11.6	36.2	60.8
0.9	49.4	32.3	14.5	**36.4**	**61.1**
0.95	46.5	**32.8**	**16.4**	35.8	60.7

Table 9: Varying γ for SEQL with λ of 4.3×10^{-4}. The accuracy is reported on ImageNet-LT `test` set. When γ is 0, the SEQL degenerates to softmax loss function.

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]

for the rarest category. CIFAR-100-LT shares the same test set of 10k images with original CIFAR-100. We report the top1 and top5 accuray. ImageNet-LT [30] is generated from ImageNet-2012 [7], which contains 1000 categories with images number ranging from 1280 to 5 images for each category [2]. There are 116k images for training and 50k images for testing. Different from CIFAR-100-LT, we additionally present accuracies of many shot, medium shot and few shot to measure the improvement on tail classes.

Implementation Details. For CIFAR-100-LT, we use Nesterov SGD with momentum 0.9 and weight decay 0.0001 for training. We use a total mini-batch size of 256 with 128 images per GPU. The model ResNet-32 is trained for 12.8K iterations with learning rate 0.2, which is then decayed by a factor of 0.1 at 6.4K and 9.6K iteration. Learning rate is increased gradually from 0.1 to 0.2 during the first 400 iterations. For data augmentation, we first follow the same setting as [24, 18], then use autoAugment [4] and Cutout [8]. In testing, we simply use the origin 32 × 32 images. For ImageNet-LT, we use a total mini-batch size of 1024 with 16 GPUs. We use ResNet-10 as our backbone like [30].The model is trained for 12K iterations with learning rate 0.4, which is divided by 10 at 3.4K, 6.8K, 10.2K iterations. A gradually warmup strategy [14] is also adopted to increase the learning rate from 0.1 to 0.4 during the first 500 iterations. We use random-resize-crop, color jitter and horizontal flipping as data augmentation. Training input size is 224 × 224. In testing, we resized the images to 256 × 256 then cropped a single view of 224 × 224 at the center.

Results on CIFAR-100-LT and ImageNet-LT. We build a much stronger baseline on CIFAR-100-LT due to those augmentation techniques. As shown in Table 7, our EQL still improves the strong baseline by a large margin of 2%. And those improvements are come from classes with fewer training samples. As for ImageNet-LT, we also present ablation studies in Table 9. A wide range of values of γ give consistent improvements over the softmanx loss baseline. As shown in Table 8 and Table 10, our equalization loss surpasses prior state-of-the-art approaches significantly, which demonstrates that our method can be generalized to different tasks and datasets effectively.

7. Conclusion

In this work, we analyze the severe inter-class competition problem in long-tailed datasets. We propose a novel equalization loss function to alleviate the effect of the overwhelmed discouraging gradients on tail categories. Our method is simple but effective, bringing a significant improvement over different frameworks and network architectures on challenging long-tailed object detection and image classification datasets.

Acknowledgment Wanli Ouyang is supported by the Australian Research Council Grant DP200103223.

[2] `https://github.com/zhmiao/OpenLongTailRecognition-OLTR`

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[6] 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.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[9] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[11] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[15] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[16] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[17] 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.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] 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.

[20] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[21] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019.

[22] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[23] 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.

[24] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.

[25] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8577–8584, 2019.

[26] 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.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[28] 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 *ECCV*, 2014.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[31] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2019.

[32] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[34] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[35] 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.

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[38] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016.

[39] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[40] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv preprint arXiv:1902.07379*, 2019.

[41] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[42] 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.

[43] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2019.

[44] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.