Long-tailed Instance Segmentation using Gumbel Optimized Loss

Konstantinos Panagiotis Alexandridis[1,2] ✉, Jiankang Deng[3], Anh Nguyen[2], and Shan Luo[1,2]

[1] King's College London, London WC2R 2LS, United Kingdom
{konstantinos.alexandridis,shan.luo}@kcl.ac.uk
[2] University of Liverpool, Liverpool L69 3BX, United Kingdom
{konsa15,anguyen,shan.luo}@liverpool.ac.uk
[3] Imperial College London, London SW7 2AZ, United Kingdom
j.deng16@imperial.ac.uk

Abstract. Major advancements have been made in the field of object detection and segmentation recently. However, when it comes to rare categories, the state-of-the-art methods fail to detect them, resulting in a significant performance gap between rare and frequent categories. In this paper, we identify that Sigmoid or Softmax functions used in deep detectors are a major reason for low performance and are suboptimal for long-tailed detection and segmentation. To address this, we develop a Gumbel Optimized Loss (GOL), for long-tailed detection and segmentation. It aligns with the Gumbel distribution of rare classes in imbalanced datasets, considering the fact that most classes in long-tailed detection have low expected probability. The proposed GOL significantly outperforms the best state-of-the-art method by 1.1% on AP, and boosts the overall segmentation by 9.0% and detection by 8.0%, particularly improving detection of rare classes by 20.3%, compared to Mask-RCNN, on LVIS dataset. Code available at: https://github.com/kostas1515/GOL.

Keywords: Long-tailed distribution, long-tailed instance segmentation, Gumbel activation

1 Introduction

There have been astonishing advancements in the fields of image classification, object detection and segmentation recently. They have been made possible by using curated and balanced datasets, e.g., CIFAR [20], ImageNet [9] and COCO [22] and by using deep Convolutional Neural Networks (CNNs). Despite that, all these advancements could be in vain if they are not usable in real-world applications. For example, the performance of classifiers in ImageNet is similar to humans, however, ImageNet pretrained detectors still struggle as they suffer from various sources of imbalance [27]. Moreover, existing instance segmentation models [12,2,5] fail to generalize for long-tailed datasets and their performance
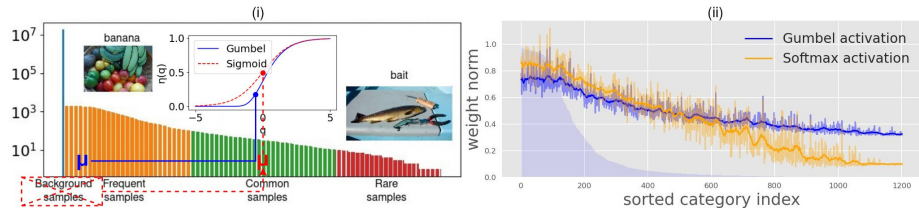
Fig. 1. (i) Gumbel activation function (blue) is asymmetric and it aligns better with the long-tailed instance segmentation distribution due to the extreme background and foreground class imbalance, whereas Sigmoid activation (red) is symmetric and more appropriate for balanced distributions. (ii) Gumbel activation (blue) produces more balanced weight norms in comparison to Softmax activation (orange) in the LVIS [11] dataset using Mask-RCNN [12].

significantly decreases for the rare categories [11]. As a result, it is difficult to exploit the advancements in image classification and transfer them to applications like object detection and segmentation due to the imbalance problem. Furthermore, there is a significant gap in performance between frequent (head) and infrequent (tail) classes in long-tailed datasets, as the state-of-the-art (SOTA) methods only detect the frequent classes [36,21]. All such problems may deteriorate the reliability of autonomous systems that rely on object detection and segmentation and raise concerns.

One possible solution for improving the long-tailed instance segmentation performance is to gather more samples for rare classes, as it is known that CNNs can achieve better results by using more data. Unfortunately, data collection will be not only costly but also intractable. The physical world contains objects that follow the Zipfian distribution [23]. This means that by increasing the distinct classes of a dataset, it is unavoidable that some will be frequent while others will be rare.

The main reason for the low performance of instance segmentation in long-tailed datasets is class imbalance. As discussed in [33,35,24], head classes dominate during training and they cause large discouraging gradients for tail classes. Since tail classes have fewer training samples, the amount of positive feedback is scarce and in the end, the model is trained effectively only for the head classes. It is also reflected by the norms of the classifier's weights [16]: classifiers trained under the long-tailed paradigm have classification weights whose norms are larger for head classes and lower for tail classes. As larger weights produce larger probabilities, the classifiers are therefore biased towards head classes. For these reasons, many prior works focus on balancing either the weight norms or the gradients caused by head and tail classes or performing two-stage training where the model is first trained for all classes and then fine-tuned for tail categories.

In contrast, we argue that the low performance in long-tailed instance segmentation is partially due to the use of sub-optimal activation functions in bounding box classification. Most classes in this long-tailed distribution have extremely low expected probabilities due to imbalance [27], making the widely
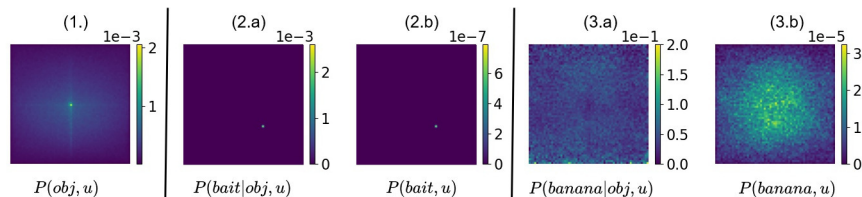
Fig. 2. Object distributions in the LVIS long-tailed object detection dataset [11]. (1): The distribution of objects $P(obj, u)$ in the dataset (irrespective of their class); (2.a): the class probability conditioned on object and its location $P(y|obj, u)$, and (2.b): the expected class distribution $P(y, u)$, for the tail class *bait*; (3.a): the class probability conditioned on object and its location $P(y|obj, u)$, and (3.b): the expected class distribution $P(y, u)$, for the head class *banana*. As shown in the figures, the distributions of objects in a long-tailed object dataset have a normal distribution as a whole and also for the head classes, whereas follows a Gumbel distribution for tail classes.

used activation functions such as Sigmoid and Softmax unsuitable. For this reason, we develop a new activation function, namely Gumbel activation and a new Gumbel loss function to model the long-tailed distribution in instance segmentation. Gumbel activation is an asymmetric function that aligns better with the long-tailed instance segmentation distribution as shown in Figure 1(i). Moreover, Gumbel loss allows the gradient of positive samples to grow exponentially while suppressing the gradient of negative samples. This is especially useful for rare category learning, in which positive feedback is scarce. At the same time, it produces more balanced classification weight norms in comparison to Softmax as shown in Figure 1(ii), suppressing the classification bias. Both head and tail categories can benefit from Gumbel loss, without the need of gradient re-balancing, exhausting parameter tuning, weight normalization or complex two-stage training. Furthermore, Gumbel is agnostic to frameworks, it can be used alongside with other loss functions and datasets, which makes it widely applicable. Based on the proposed Gumbel loss, we have developed Gumbel optimized methods, that outperform the state-of-the-art instance segmentation methods on the LVIS [11] dataset. We list our contributions as follows:

- We identify the problem of activation functions in long-tailed instance segmentation for the first time, via extensive experiments;
- We propose a new loss, i.e., Gumbel Optimized Loss (GOL), for long-tailed instance segmentation;
- We have validated the effectiveness of GOL on real-world long-tailed instance segmentation datasets, outperforming the SOTA methods by a large margin.

2   Related works

Long-tailed object classification. It has been a hot topic to address the imbalance problem in object classification. Long-tailed object classification datasets

of CIFAR10, CIFAR100 and ImageNet have been investigated to tackle imbalanced classification using techniques such as data re-sampling [4,24,31,42], Cost Sensitive Learning (CSL) [8,17], margin adjustment [25,18,3,30] and two stage training [16,36]. Data re-sampling methods re-sample the rare classes and have been most widely used and investigated [4,24,31,42]. However, such methods cost more training effort and pose the risk of overfitting for rare classes, while under-sampling under-fits heads class and deteriorates the overall performance. The CSL methods construct a cost matrix so that the cost function can be more sensitive to the rare classes [8,17], so as to exploit the data available. But CSL methods are dependent on the dataset and require careful calibration to avoid exploding gradients caused by excessive costs. Margin adjustment techniques change the decision boundary of the classifiers by either normalizing the classifier weight norms, engineering appropriate losses or modifying the classification prediction a-posteriori [25,18,3,30], which do not cost additional training time. Their drawback is that the margins are difficult to compute, and they are based on dataset statistics.

Long-tailed object detection. Some methods addressing the long-tailed image classification could be applied in long-tailed object detection [30,33,35,16]. However, many SOTA long-tailed classification methods obtain low performance for tasks that include the special *background* class [26]. Under this realistic scenario, the performance drop is caused by the extreme imbalance between the dominant *background* class and other foreground classes. The same applies in long-tailed object detection where the imbalance factor is ~ 1000 larger than the imbalance factor in image classification. For this reason, not all long-tailed classification methods are transferable to long-tailed instance segmentation. Instead, many methods are developed to tackle long-tailed object detection, directly. Some of them include the creation of specialized loss functions that balance the gradient contribution of positive and negative samples [33,32,37,35,29]. Others construct hierarchical groups [21,38], enforce margins in the classifier's prediction [10,30,34] or use two-stage strategies [36,16,40]. These methods have produced promising results but they suffer from limitations: Two-stage methods are complex and laborious; hierarchical methods require pre-processing and careful grouping; loss engineering methods have many hyper-parameters that need tuning. All these methods use Sigmoid or Softmax as their activation function, which is not close to the target distribution and not a good choice as we discuss in Section 3.

To the best of our knowledge, we are the first to tackle long-tailed segmentation by using Gumbel loss function. The most related work is [1], where they used the general extreme value distribution to classify Covid-19 cases. In contrast, we develop Gumbel for long-tailed instance segmentation.

3   Problem Formulation

Assume a dataset $X = \{x_i, y_i\}, i \in \{1, ..., N\}$, where x_i, y_i are images and annotations respectively and N is the total number of training images. We can train

a convolutional neural network $f(X, \theta) = z$, where z is the latent representation and θ is the network's weight parameters. To calculate the prediction $\bar{y}$, one first can use a fully connected layer $q(z) = W^T z + b$, where W is the classification weights and b is the bias term, to calculate the score q_i. Then $\bar{p}_i = \eta(q_i)$ is used to transform the score q_i into probability $\bar{p}_i$, using the activation function $\eta(\cdot)$ and finally the prediction $\bar{y}$ is calculated using $\bar{y} = \arg\max_i(\bar{p}_i)$.

In image classification and instance segmentation, the Sigmoid activation, i.e., $\eta_{\text{sigmoid}}(q_i) = \frac{1}{1+e^{-q_i}}$, or the Softmax activation, i.e., $\eta_{\text{softmax}}(q_i) = \frac{e^{q_i}}{\sum e^{q_j}}$, has been commonly used. For the binary case, it assumes that $\bar{p}_i$ follows a Bernoulli distribution as the score $q_i = \log\frac{\bar{p}_i}{1-\bar{p}_i}$ and it can be interpreted as the odds-ratio of the event $\bar{p}_i$, i.e., how many times an event happens $\bar{p}_i$ divided by how many times it does not happen $1 - \bar{p}_i$ in a log scale.

It would be reasonable to use the Sigmoid or Softmax activation function for image classification, where the expected probability distribution $\mathbb{P}$ is a Bernoulli distribution and all classes are mutually exclusive, thus one can use Sigmoid or Softmax to effectively model the data. However, we argue that it would not be well suitable to use these activation functions for long-tailed instance segmentation as the expected distribution of objects is not the same as the expected image distribution in classification. Object distribution is more complex as it is affected by class imbalance and location imbalance.

3.1   Object Distribution

To make this clear, we first calculate the ground truth object distribution. To this end, we calculate the expected number of objects $P(obj, u)$ whose centers fall inside the cell $u = [i, j]$ of the normalized grid as follows:

$$P(obj, u) = \frac{\sum_{x=1}^{x=N} \mathbb{1}(obj, x)\mathbb{1}(obj, u)}{M} \tag{1}$$

where obj is the object occurrence, $\mathbb{1}$ is the indicator function and M is the total number of objects in the dataset. Next, we calculate the class membership $P(y|obj, u)$ for each location u, which summarizes the uncertainty of an object belonging to each class y in the dataset for the specific location u (i.e., it holds that $\sum_{y=1}^{y=C} P(y|obj, u) = 1$). Finally, we calculate $P(y, u)^4$ as:

$$P(y, u) = P(y|obj, u)P(obj, u) \tag{2}$$

The final target distribution is a distribution that we aim to estimate by minimizing the cross entropy between the target and the data distribution.

In Figure 2, we visualize $P(obj, u)$, $P(y|obj, u)$ and $P(y, u)$ for one head class and one tail class of LVIS dataset. The probability of detection for a head class i.e., *banana* in the example, is low in each location and varies in different locations of the image. For a tail class, i.e., *bait* in the example, is even lower and zero for

---

[4] Here we omit obj for simplicity since $P(y, obj, u) = P(y, u)$ due to that y shows there is object occurrence obj.

most locations. This is different from long-tailed image classification, in which the expected class probabilities are not affected by location imbalance, only by class imbalance. On the other hand, in long-tailed instance segmentation, classes have even lower expected probabilities as they are affected by both location imbalance and class imbalance. For example, even head classes like *banana* have even lower expectation for locations far away from the center of the image, making the long-tailed segmentation more challenging for both head and tail classes. This highlights the magnitude of imbalance in long-tailed segmentation and motivates us to develop Gumbel activation.

Using Gumbel, we assume that the target distribution follows Gumbel distribution and this is a better choice than using Sigmoid or Softmax because the expected classification probabilities are minuscule. In fact, by using any activation function, one assumes how the ground truth is distributed. It is a common practice to use Sigmoid or Softmax and this assumes that the target distribution is Bernoulli. While this is a rational choice for image classification, it is unrealistic for long-tailed instance segmentation as the expected classification probabilities are infinitesimal. For this reason, we assume that the target distribution is Gumbel and we use Gumbel activation. We further explain why choosing an activation, that implicitly assumes the target distribution, below.

3.2   Activations as Priors

To understand why choosing an activation implicitly assumes the target distribution, we consider an example of binary classification, however, it can be extended to multi-class classification easily. In a binary classification problem, the true variable y relates with the representations z as follows:

$$y = \begin{cases} 1, & \text{if } W^T z + b + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where ϵ is the error that is a random variable. The classification boundary is set to 0, but it could be any other value as it is adjusted by the bias term b in optimization. We are interested in the probability $P(y = 1)$ and this is calculated as:

$$\begin{aligned} P(y = 1) &= P(W^T z + b + \epsilon > 0) \\ P(\epsilon > -W^T z - b) &= 1 - F(-q) \end{aligned} \tag{4}$$

where F is the cumulative distribution function. Many practitioners use Sigmoid to activate q and estimate $P(y = 1)$ and this means that:

$$\begin{aligned} P(y = 1) = \eta_{\text{sigmoid}}(q) &= \frac{1}{1 + e^{-q}} \\ \eta_{\text{sigmoid}}(q) &= F_{\text{logistic}}(q; 0, 1) \\ P(y = 1) &= 1 - F_{\text{logistic}}(-q; 0, 1) \end{aligned} \tag{5}$$

By comparing Eq. 4 and the last expression of Eq. 5, it is understood that by using Sigmoid activation, one assumes that the error term ϵ follows the standard

logistic distribution (with $\mu = 0$ and $\sigma = 1$), as one chooses F to be logistic. In practice, when any activation function $\eta(q)$ is applied, it is implicitly assumed how ϵ is distributed and as a result the target distribution $y \sim \mathbb{P}$ is assumed. For example, if a Sigmoid function is used, the variable q is transformed into a binomial probability distribution, which implies that it has a success rate $\bar{p}$ and a failure rate $1 - \bar{p}$, like a coin toss. In this case, it is assumed that y follows Bernoulli Distribution and the error ϵ follows Logistic Distribution. Finally, the training of the model is to minimize the discrepancy between the target distribution $\mathbb{P}$ and the predicted distribution $\mathbb{Q}$ using Cross Entropy:

$$H(\mathbb{P}(x), \mathbb{Q}(x)) = - \sum_{x \in X} \mathbb{P}(x) \log(\mathbb{Q}(x)) \tag{6}$$

In training, as Stochastic Gradient Descent is an iterative algorithm, the starting conditions play a significant role. This suggests that it is preferable to have a good starting prior so that the initial estimation of $\mathbb{P}$ is reasonable and the choice of the activation function will facilitate the initial estimation. This has a similar concept to the prior distribution in Bayesian Inference, where it is important to choose a suitable prior for optimal results.

Hypothesis. In long-tailed instance segmentation, the expected classification probabilities are significantly low due to imbalance problems as mentioned by [27] and explained in Eq. 2. For this reason, we hypothesize that Gumbel activation will produce superior results as it models the data using Gumbel Distribution which is closer to the real object distribution.

In conclusion, long-tailed instance segmentation is far more challenging than image classification and naive usage of Sigmoid or Softmax activation infringes on the underlying assumptions. Nevertheless, one can use Cross Entropy to minimize the discrepancy between the target distribution $\mathbb{P}$ and the predicted distribution $\mathbb{Q}$ but if the prior guess is significantly different than the actual distribution $\mathbb{P}$ then the results might be sub-optimal. By choosing the activation function, one guesses how the target $\mathbb{P}$ is distributed. For Sigmoid or Softmax, one believes that $\mathbb{P}$ is Bernoulli distribution and while this is reasonable for image classification, we argue that in long-tailed instance segmentation it is not optimal and we empirically show that Gumbel can produce superior results.

4 Gumbel Activation for Long-tailed Detection

4.1 Sigmoid Activation for Object Classification

It is useful to remind the readers about Sigmoid activation as we can make clear distinctions between this and our suggested activation. The formula is $\eta_\sigma(q_i) = \frac{1}{1 + \exp(-q_i)}$. If one encodes the ground truth y as a one-hot vector then the gradient using Eq. 6 is $\frac{dL(\eta_\sigma(q_i), y_i)}{dq_i} = y_i(\eta_\sigma(q_i) - 1) + (1 - y_i)\eta_\sigma(q_i)$.

Sigmoid is a symmetric activation function: the positive gradient (i.e., when $y = 1$) takes values from $(-1, 0)$, while the negative gradient (i.e., when $y = 0$) takes values from $(0, 1)$ and the response value grows with the same rate for both positive and negative input, as shown in Figure 3(i).

4.2   Gumbel Activation for Rare-class Segmentation

We notice that $P(y, u)$ has infinitesimal probabilities. For tail classes it has a maximum value in the scale of *1e-7*, for head classes it has a maximum *1e-5* and for both cases, it has even lower probabilities for edge locations in the image. This motivates us to model $P(y, u)$ using the extreme value distribution Gumbel. Gumbel is useful for modelling extreme events, i.e., those with very low probabilities. For example, Gumbel can be used to predict the next great earthquake because this has a much lower probability than regular earthquakes. Other applications of Gumbel can be found in finance and biology and the readers are referred to this work [19] for more information on extreme value distribution. In long-tailed instance segmentation, one can use the cumulative density function of the standard Gumbel distribution as the activation function (we further study the choice of non-standard Gumbel activation in supplementary material):

$$\eta_\gamma(q_i) = F_{\text{Gumbel}}(q; 0, 1) = \exp(-\exp(-q_i)) \tag{7}$$

Combining Eq. 6 and Gumbel activation we derive Gumbel Loss (GL) as:

$$GL(\eta_\gamma(q_i), y_i) = \begin{cases} -\log(\eta_\gamma(q_i)), & if \;\; y_i = 1 \\ -\log(1 - \eta_\gamma(q_i)), & if \;\; y_i = 0 \end{cases} \tag{8}$$

The gradient of Eq. 8 is:

$$\frac{dL(\eta_\gamma(q_i), y_i)}{dq_i} = \begin{cases} -\exp(-q_i), & if \;\; y_i = 1 \\ \frac{\exp(-q_i)}{\exp(\exp(-q_i)) - 1}, & if \;\; y_i = 0 \end{cases} \tag{9}$$

The Gumbel activation is asymmetric as illustrated in Figure 3(i). This means that the positive gradients (i.e., when $y = 1$) will take values from $(-\infty, 0)$ while the negative gradients (i.e., when $y = 0$) will take values from $(0, 1)$. This is a beneficial property that allows the positive feedback to grow exponentially while it suppresses the negative feedback. It is especially useful in long-tailed segmentation as the positive feedback for rare categories is scarce. With Gumbel activation, positive gradients can grow faster than by using Sigmoid activation as shown in Figure 3(ii) and this can boost the performance of rare classes.

4.3   Gumbel Optimised Loss

Gumbel activation can be used not only with Cross Entropy loss but with other state-of-the-art loss functions as well. We select the recently proposed Droploss [14] as the loss function of our enhanced Gumbel-based model, other loss functions are shown in Fig. 4(ii). Using Gumbel activation and Droploss we propose Gumbel Optimised Loss GOL as follows:

$$\mathcal{L}_{GOL} = -\sum_{j=1}^{C} \log(w_j^{Drop} \bar{p}_j), \;\; \bar{p}_j = \begin{cases} \eta_\gamma(q_i), & if \;\; y_j = 1 \\ 1 - \eta_\gamma(q_i), & if \;\; y_j = 0 \end{cases} \tag{10}$$

where w_j^{Drop} are class specific weights proposed by DropLoss [14]. We show the full equation in supplementary material due to space limitations.
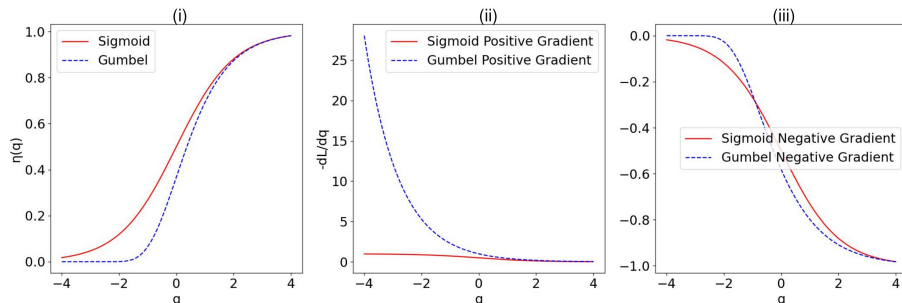
Fig. 3. Activation properties for Gumbel (blue dashed lines), Sigmoid (red solid lines) activation functions. Their activation behaviours are illustrated in (i). The Gumbel activation is asymmetric, whereas Sigmoid is symmetric. Their positive gradients and negative gradients are illustrated in (ii) and (iii) respectively, using inverted y-axis $\frac{-dL}{dq}$. The positive gradients of the Gumbel activation (i.e., when $y = 1$) ranges in $(-\infty, 0)$, while the negative gradients (i.e., when $y = 0$) ranges in $(0, 1)$.

5 Experimental Setup

Dataset and evaluation metrics. For our experiments on long-tailed instance segmentation, we use LVIS (Large Vocabulary Instance Segmentation) dataset [11]. We mainly use version 1 which contains 100k images for training and 19.8k images for validation. LVISv1 dataset contains $1,203$ categories that are grouped according to their image frequency: *rare* categories (those with 1-10 images in the dataset), *common* categories (11 to 100 images in the dataset) and *frequent* categories (those with > 100 images in the dataset.) Some previous methods use the LVISv0.5 dataset, which has 1230 classes instead of 1203. For fairness, we also show results in this dataset. We report our results using average segmentation performance AP, average box performance AP^b and average segmentation performance for rare AP^r, common AP^c and frequent categories AP^f.

Implementation details. For our experiments, we use 4 V100 GPUs, a total batch size of 16 images, a learning rate of 0.02, weight decay of 0.0001, Stochastic Gradient Descent and momentum of 0.9. We use random horizontal flipping and multi-scaling as data augmentations for training, following the conventions of the community. We train our models using the *mmdetection* framework [6] and during inference, we use score threshold of 0.0001 as in [11].

For our intermediate experiments, we use Gumbel activation and a plethora of architectures, backbones, loss functions and sampling strategies using the 1x schedule. For our enhanced GOL model, we use Mask-RCNN [12], the 2x schedule, Normalised Mask [35], RFS sampler [11] and a stricter Non-Maximum Suppression threshold that is 0.3. When using Gumbel activation, we initialize the weights of the classifier to 0.001 and the bias terms to -2.0, to enable stable training. More implementation details, design choices and results are discussed in our supplementary material.

Table 1. Comparative results for LVISv1 using schedule 1x, random sampler (left) and RFS [11] sampler (right). Gumbel activation is superior than Softmax, especially for the case of random sampling when the distribution is unaltered.

Sampler	Random					RFS [11]				
Method	AP	AP^r	AP^c	AP^f	AP^b	Method	AP	AP^r	AP^c	AP^f	AP^b
Sigmoid	16.4	0.8	12.7	27.3	17.2	Sigmoid	22.0	11.4	20.9	27.9	**23.0**
Softmax	15.2	0.0	10.6	26.9	16.1	Softmax	21.5	9.7	20.7	27.6	22.4
Gumbel (ours)	**19.0**	**4.9**	**16.8**	**27.6**	**19.1**	Gumbel (ours)	**22.7**	**12.2**	**21.2**	**28.0**	22.9

6   Results

6.1   Results with Different Sampling Strategies

We compare activations when using a random sampler and 1x schedule. Under this configuration, the target distribution is unaltered and the probability of sampling is equal to the dataset's class probability. As Table 1 suggests, the best activation function using a random sampler is Gumbel which largely outperforms Sigmoid and Softmax. It increases overall AP by 2.6%, AP^r and AP^c by 4.1% and AP^b by 1.9% compared to Sigmoid and AP by 3.8%, AP^r by 4.9%, AP^c by 6.2 % and AP^b by 3.0% compared to Softmax. Noticeably, the gap in mask and box performance is smaller with Gumbel activation at 19.0% and 19.1% respectively, while other activations have larger gaps between box and segmentation performance. This suggests that Gumbel is more suitable for long-tailed segmentation than other activation functions.

   We apply the state-of-the-art RFS [11] method, which is an oversampling method and report the results in Table 1. With RFS, images containing rare categories are up-sampled, thus the original distribution is distorted. Under this scenario, Gumbel activation does not boost the performance as much as before, as the object distribution becomes more balanced with oversampling. Nevertheless, Gumbel improves overall performance by a respective margin of 0.7% in overall segmentation and by 0.8% AP^r which is attributed to the fact that firstly, there are still classes in the dataset that suffer from imbalance after using RFS, thus Gumbel can model them better than Sigmoid or Softmax; secondly, Gumbel activation has a lower gap in bounding box and segmentation performance than Sigmoid or Softmax, which results in higher segmentation performance (i.e., 0.7% increase) given similar box performance.

6.2   Integrating Gumbel Activation

We conduct experiments using training schedule 1x with larger backbones, i.e., Resnet101 [13] and ResNeXt101 [39] and architectures i.e., Cascade Mask RCNN [2] and Hybrid Task Cascade [5] to determine if the proposed activation generalizes for deeper models. As shown in Figure 4 (i), Gumbel activation is a better choice than the Softmax activation function as it achieves better overall AP and AP in rare categories.
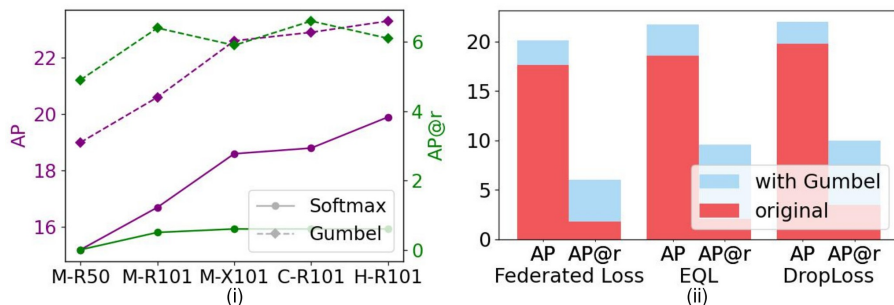
Fig. 4. (i): Comparison of Softmax against Gumbel activation using common instance segmentation frameworks. (ii): Comparison of Sigmoid-based SOTA losses against the Gumbel-based alternatives. All models use 1x schedule and random sampling.

Next, we examine the behavior of the Gumbel activation function using SOTA loss functions Equalization loss (EQL) [33], DropLoss [14] and Federated Loss [41]. For EQL and DropLoss, we use the hyperparameter $\lambda = 0.0011$ which is more appropriate for LVISv1 as described in [32] and we change only the activation function from Sigmoid to Gumbel. For Federated Loss, we use the same hyper-parameters as described in [41] changing only the activation from Sigmoid to Gumbel. As Figure 4 (ii) indicates, Gumbel significantly boosts the performance of all models in both overall AP and rare category AP and this highlights its applicability and efficacy. We show more detailed results in our supplementary material.

6.3 GOL Components

We conduct an ablation study using a 2x-schedule and we report the most significant findings, a more detailed ablation study is provided in our supplementary material. We use the standard Mask-RCNN, EQL [33] and RFS [11] as the basis and examine the behavior of Gumbel. As shown in Table 2, Gumbel can significantly boost the performance of this pipeline by 0.8%. To further boost the performance, we use a stricter Non-Maximum Suppression threshold that is 0.3 and Mask normalization, we denote these enhancements in the Table as (Enh). Next, we adopt DropLoss which is a recent improvement in the loss function of EQL, proposed by [14]. Finally, the best performance is achieved using RFS, Gumbel, Enh and DropLoss, we codename this as pipeline *GOL*.

Total performance. In the end, our *GOL* method significantly improves the vanilla Mask-RCNN *AP* by 9.0%, and it largely improves AP^r by 20.3%, AP^c by 11.5%, AP^f by 1.2% and AP^b by 8.0%.

Comparison with other methods. As shown in Table 3, *GOL* significantly surpasses the state-of-the-art in LVIS Dataset using the standard Mask-RCNN benchmark. In detail, in LVISv0.5, *GOL* achieves 29.5% *AP*, surpassing RFS [11] by 4.1% and the best state-of-art LOCE [10] by 1.1% in *AP*. Moreover,

Table 2. Ablation study, using Mask-RCNN, Resnet50 and training schedule 2x.

RFS	EQL	Gumbel	Enh	DropLoss	AP	AP^r	AP^c	AP^f	AP^b
					18.7	1.1	16.2	29.2	19.5
✓					23.7	13.3	23.0	29.0	24.7
✓	✓				25.3	17.4	24.9	29.2	26.0
✓	✓	✓			26.1	18.4	25.9	29.8	26.8
✓	✓	✓	✓		26.9	18.1	26.5	**31.3**	26.8
✓		✓	✓	✓	**27.7**	**21.4**	**27.7**	30.4	**27.5**
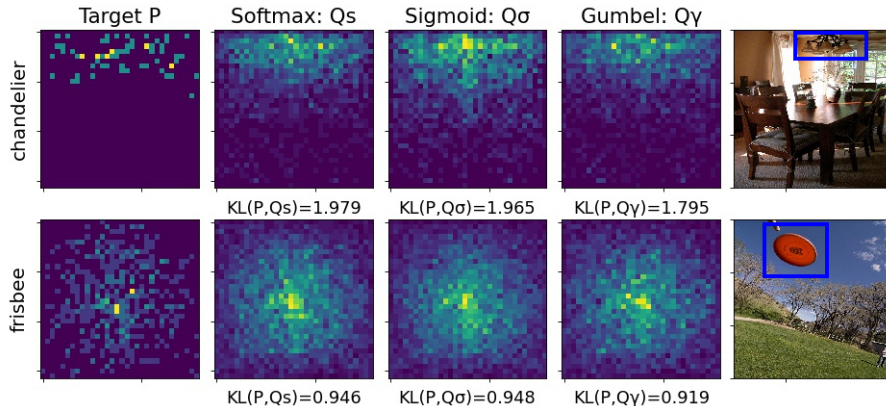
Fig. 5. Comparison of two object distributions in LVIS validation set, using Softmax (second column), Sigmoid (third column) and Gumbel (fourth column). Gumbel predicts distributions that have smaller KL divergence than Sigmoid or Softmax.

our method achieves the best performance at rare and common categories and it consistently surpasses many other recent works. In LVISv1, *GOL* achieves 27.7% overall AP, surpassing RFS[11] by 4.0%, LOCE by 1.1%, Seesaw Loss [35] by 1.3% and EQLv2[32] by 2.2% using ResNet50 backbone. It also achieves the best AP, which is at least 0.9% higher than other methods, using the larger ResNet101 backbone. Finally, it consistently outperforms all other methods for rare and common categories in both Resnet50 and Resnet101 backbones.

6.4   Model Analysis

We analyze the behavior of Mask-RCNN using Gumbel Loss. In detail, we visualize Mask-RCNN predicted object distributions in the validation set, for two random classes *chandelier* and *frisbee*. We compare, the predicted distributions $\mathbb{Q}$ of Softmax, Sigmoid and Gumbel against the ground truth $\mathbb{P}$ using the Kullback Leibler divergence.

As Figure 5 suggests, Gumbel produces object distributions that are closer to the target distribution, as they have smaller Kullback–Leibler (KL) divergence.

Table 3. Comparison against SOTA on the LVIS dataset.

Method	Dataset	Framework	AP	AP^r	AP^c	AP^f	AP^b
RFS [11]			25.4	16.3	25.7	28.7	25.4
DropLoss[14]			26.4	17.3	28.7	27.2	25.8
BAGS [21]			26.2	18.0	26.9	28.7	25.8
BALMS[30]			27.0	19.6	28.9	27.5	27.6
Forest-RCNN[38]	LVIS v0.5	Mask-RCNN R50-FPN	25.6	18.3	26.4	27.6	25.9
EQLv2[32]			27.1	18.6	27.6	29.9	27.0
LOCE[10]			28.4	22.0	29.0	30.2	**28.2**
DisAlign [40]			27.9	16.2	29.3	**30.8**	27.6
GOL (ours)			**29.5**	**22.5**	**31.3**	30.1	**28.2**
RFS[11]			23.7	13.3	23.0	29.0	24.7
EQLv2[32]			25.5	17.7	24.3	30.2	26.1
LOCE[10]	LVIS v1.0	Mask-RCNN R50-FPN	26.6	18.5	26.2	**30.7**	27.4
NorCal with RFS [28]			25.2	19.3	24.2	29.0	26.1
Seesaw[35]			26.4	19.5	26.1	29.7	**27.6**
GOL (ours)			**27.7**	**21.4**	**27.7**	30.4	27.5
RFS [11]			25.7	17.5	24.6	30.6	27.0
EQLv2[32]			27.2	20.6	25.9	31.4	27.9
LOCE[10]	LVIS v1.0	Mask-RCNN R101-FPN	28.0	19.5	27.8	**32.0**	29.0
NorCal with RFS [28]			27.3	20.8	26.5	31.0	28.1
Seesaw[35]			28.1	20.0	28.0	31.8	28.9
GOL(ours)			**29.0**	**22.8**	**29.0**	31.7	**29.2**

Moreover, as [16] has suggested, there is a positive correlation between the weight norms of the classifier and the image frequency of categories, which results in classification bias. In our case, we visualize the weight norms of the Mask-RCNN classifier trained with Softmax (baseline model), and Gumbel respectively. As Figure 1 (ii) suggests, the weight norm distribution of the Mask-RCNN classifier trained with Gumbel is more uniform than the distribution of vanilla Mask-RCNN. This suggests that the classifier norms are more balanced when using Gumbel loss which validates its efficacy.

6.5   Long-tailed Image Classification

We further test Gumbel activation in long-tailed image classification benchmarks. For all classification experiments, we use random sampling and decoupled strategy. In decoupled strategy, the model is first trained with Softmax activation and then only the classifier is re-trained with Gumbel activation.

CIFAR100-LT [3]. We train a ResNet32 for 240 epochs using Auto-Augment [7], SGD, weight decay 0.0002, batch size 64 and learning rate 0.1 that decays at epoch 200 and 220 by 0.01. In the second stage, we retrain the classifier using a learning rate of 1e-4 for 15 epochs.

ImageNet-LT, Places-LT [23]. For ImageNet-LT, we train ResNet50 with and without Squeeze and Excite [15] modules for 200 epochs using Auto-Augment. For Places-LT, we finetune an ImageNet pretrained ResNet152 for 30 epochs.

For both datasets, we use SGD, weight decay 0.0005, batch size 256, and learning rate 0.2 with cosine scheduler. In the second stage, we retrain the classifier using a learning rate of 1e-5 for 10 epochs.

Table 4. Top-1 accuracy on long-tailed classification datasets.

Dataset	CIFAR100-LT			ImageNet-LT		Places-LT
Imbalance factor	50	100	200	256		996
Model		ResNet-32		ResNet-50	SE-ResNet-50	ResNet-152
Softmax	46.2	42.4	38.3	45.2	45.9	28.7
Gumbel	**49.0**	**45.5**	**41.5**	**48.2**	**48.5**	**30.0**

As Table 4 indicates, Gumbel activation can boost the classification performance of all models in all datasets consistently.

7   Conclusion and Discussions

We hypothesize that real-world long-tailed detection and segmentation data follows a distribution that is closer to Gumbel distribution and not Bernoulli. For this reason, we propose to use Gumbel activation instead of Sigmoid or Softmax. We validate the superiority of Gumbel against Sigmoid and Softmax under different sampling strategies, deeper models and loss functions and we develop the *GOL* method based on Gumbel activation that significantly outperforms the state-of-the-art. Our extensive experiments validate that Gumbel is a superior activation function that can be used as a component with both off-the-shelf methods and state-of-the-art models to further increase their performance.

We have also tested Gumbel activation in long-tailed classification benchmarks and saw consistent improvements when Gumbel is used as a decoupled method. Finally, Gumbel could also be used for dense object detection and we have seen a 0.4% increase in AP when using RetinaNet on COCO and 1x schedule. Currently, Gumbel cannot be used with Softmax-based loss functions and it does not take full advantage of oversampling methods. In the future, we will develop a custom loss function and sampling mechanism tailored to Gumbel activation.

References

1. Bridge, J., Meng, Y., Zhao, Y., Du, Y., Zhao, M., Sun, R., Zheng, Y.: Introducing the gev activation function for highly unbalanced data to develop covid-19 diagnostic models. IEEE journal of Biomedical and Health Informatics **24**(10), 2776–2786 (2020)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems (2019)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
5. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)
6. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
7. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
8. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Feng, C., Zhong, Y., Huang, W.: Exploring classification equilibrium in long-tailed object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3417–3426 (2021)
11. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hsieh, T.I., Robb, E., Chen, H.T., Huang, J.B.: Droploss for long-tail instance segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1549–1557 (2021)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)

16. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Eighth International Conference on Learning Representations (ICLR) (2020)
17. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems **29**(8), 3573–3587 (2017)
18. Kim, B., Kim, J.: Adjusting decision boundary for class imbalanced learning. IEEE Access **8**, 81674–81685 (2020)
19. Kotz, S., Nadarajah, S.: Extreme value distributions: theory and applications. World Scientific (2000)
20. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
21. Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10991–11000 (2020)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
23. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2537–2546 (2019)
24. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018)
25. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=37nvvqkCo5`
26. Mullapudi, R.T., Poms, F., Mark, W.R., Ramanan, D., Fatahalian, K.: Background splitting: Finding rare classes in a sea of background. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8043–8052 (2021)
27. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: A review. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3388–3415 (2020)
28. Pan, T.Y., Zhang, C., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., Chao, W.L.: On model calibration for long-tailed object detection and instance segmentation. Advances in Neural Information Processing Systems **34** (2021)
29. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., Yan, J.: Large-scale object detection in the wild from imbalanced multi-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9709–9718 (2020)
30. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. In: Proceedings of Neural Information Processing Systems(NeurIPS) (Dec 2020)
31. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: European conference on computer vision. pp. 467–482. Springer (2016)
32. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1685–1694 (2021)

33. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11662–11671 (2020)
34. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. Advances in Neural Information Processing Systems **33**, 1513–1524 (2020)
35. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9695–9704 (2021)
36. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. In: European Conference on computer vision. pp. 728–744. Springer (2020)
37. Wang, T., Zhu, Y., Zhao, C., Zeng, W., Wang, J., Tang, M.: Adaptive class suppression loss for long-tail object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3103–3112 (2021)
38. Wu, J., Song, L., Wang, T., Zhang, Q., Yuan, J.: Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1570–1578 (2020)
39. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
40. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: A unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2361–2370 (2021)
41. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. In: arXiv preprint arXiv:2103.07461 (2021)
42. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018)