阿里巴巴提出 DR Loss：解决目标检测的样本不平衡问题

点击上方“CVer”，选择加"星标"或“置顶”

重磅干货，第一时间送达

作者：张凯

https://zhuanlan.zhihu.com/p/75896297

背景

《DR Loss: Improving Object Detection by Distributional Ranking》作者来自于阿里巴巴。该论文主要是修改损失函数来处理样本不平衡问题的，之前最出名的应该是2017 ICCV最佳学生论文RetinaNet中的focal loss。2019 AAAI的GHM，2019 CVPR的AP loss也分别讨论了样本不平衡的问题。

因为这类方法只会影响训练，不会影响推理速度，对现有产品影响不会很大，所以还是很值得尝试的。

DR Loss: Improving Object Detection by Distributional Ra

Qi Qian　Lei Chen　Hao Li　Rong Jin

Alibaba Group, Bellevue, WA, 98004, USA

{qi.qian, fanjiang.cl, lihao.lh, jinrong.jr}@alibaba-inc.com

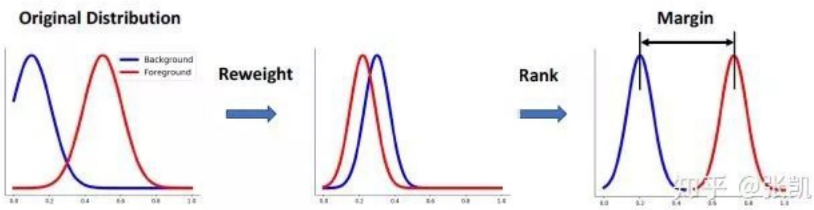
arXiv：https://arxiv.org/abs/1907.10156

代码未开源，基于detectron开发。

一、研究动机

样本不平衡问题是one-stage目标检测算法中一直存在的问题，负样本（背景）的数目远大于正样本，简单样本远大于难例，从而导致训练无法收敛到很好的解。2017 ICCV RetinaNet是通过focal loss来处理该问题，主要是抑制大量简单的负样本，给难例更大的权重。而本篇论文则提出了另外一种解决思路（2019 CVPR AP loss 也是这个思路）：将分类问题转换为排序问题，从而避免了正负样本不平衡的问题。同时针对排序，提出了排序的损失函数DR loss，并给出了可求导的解。最终性能较RetinaNet有近2个点的提升，提升还是比较明显的。

二、具体方法



整体思路的话，如图所示，主要是将正样本的分布和负样本的分布尽可能区别开，具体结合公式来讲下，比较简单。

首先是对原有分类问题的定义

$$\min_{\theta} \sum_{i}^{N} \sum_{j,k} \ell(p_{i,j,k})$$

对于所有的样本，寻找一个分类器

，使得分类损失最小，一般采用cross entropy loss，i, j, k 分别代表图像、样本、类别。

进一步地，把正负样本拆开写

$$\min_{\theta} \sum_{i}^{N} \left(\sum_{j_+}^{n_+} \ell(p_{i,j_+}) + \sum_{j_-}^{n_-} \ell(p_{i,j_-}) \right)$$

把上述问题转换为排序问题：

$$\min_{\theta} \sum_{i}^{N} \sum_{j_+}^{n_+} \sum_{j_-}^{n_-} \ell(p_{i,j_-} - p_{i,j_+} + \gamma)$$

上述公式的含义是，对于所有样本对（一个正样本和一个负样本构成一对）的损失最小，每一个样本对排序都要正确，r 代表margin。

进一步，对于每幅图像可以写成

$$\frac{1}{n_+ n_-} \sum_{j_+}^{n_+} \sum_{j_-}^{n_-} \ell(p_{i,j_-} - p_{i,j_+} + \gamma)$$
$$= E_{j_+, j_-}[\ell(p_{i,j_-} - p_{i,j_+} + \gamma)]$$

如果按照上述公式来做，会存在两个问题，一是负样本之间本身就是不平衡的，二是样本对太多了，具体是 n+ x n- 。

所以一种解决方案是改求正负样本分布的min和max：

$$\min_{\theta} \sum_{i}^{N} \ell(\max_{j_-} p_{i,j_-} - \min_{j_+} p_{i,j_+} + \gamma)$$

成功地将 量级转换为了1。但上述同样存在一个问题，就是该公式对outliers太敏感了，训练肯定不稳定。

为了解决上述问题，本文的思路是选取正负样本中最具代表性的样本来参与排序，具体地，作者定义了正样本分布和负样本分布的分数：

$$P_{i,+} = \sum_{j_+}^{n_+} q_{i,j_+} p_{i,j_+}; \quad P_{i,-} = \sum_{j_-}^{n_-} q_{i,j_-} p_{i,j_-}$$

其中q代表的是分布，并有

$\Sigma q = 1$

可以看到，如果q服从均匀分布，实际上求得就是正负样本的期望。（当然这样肯定不行，因为负样本中难易样本是不均衡的）

所以作者希望求解这个分布，使得分布的分数最小化或者最大化

$$P_{i,+} = \min_{\mathbf{q}_{i,+} \in \Delta} \sum_{j_+}^{n_+} q_{i,j_+} p_{i,j_+}; \quad P_{i,-} = \max_{\mathbf{q}_{i,-} \in \Delta} \sum_{j_-}^{n_-} q_{i,j_-} p_{i,j_-}$$

如果分布没有约束的话，那么产生的解一定是最大值对应的q为1，其余值为0，这样就又退化了之前直接求max和min了。

所以作者在此处加入了对分布的约束：

$$\begin{aligned} P_{i,-} &= \max_{\mathbf{q}_{i,-} \in \Delta, \Omega(\mathbf{q}_{i,-}) \geq \epsilon_-} \sum_{j_-}^{n_-} q_{i,j_-} p_{i,j_-} \\ -P_{i,+} &= \max_{\mathbf{q}_{i,+} \in \Delta, \Omega(\mathbf{q}_{i,+}) \geq \epsilon_+} \sum_{j_+}^{n_+} q_{i,j_+} (-p_{i,j_+}) \end{aligned}$$

进一步，转换为以下次优问题：

$$\begin{aligned} & \max_{\mathbf{q}_{i,-} \in \Delta} \sum_{j_-} q_{i,j_-} p_{i,j_-} \\ s.t. \quad & \Omega(\mathbf{q}_{i_-}) \geq \epsilon_- \end{aligned}$$

利用对偶法转换：

$$\max_{\mathbf{q}_{i,-}\in\Delta} \sum_{j-} q_{i,j_-} p_{i,j_-} + \lambda_- \Omega(\mathbf{q}_{i,-})$$

再用KKT条件，可以求得：

$$q_{i,j_-} = \frac{1}{Z_-} \exp(\frac{p_{i,j_-}}{\lambda_-}); \quad Z_- = \sum_{j_-} \exp(\frac{p_{i,j_-}}{\lambda_-})$$

最后，代入公式，求得分布的分数

$$\hat{P}_{i,-} = \sum_{j-}^{n_-} q_{i,j_-} p_{i,j_-} = \sum_{j-}^{n_-} \frac{1}{Z_-} \exp(\frac{p_{i,j_-}}{\lambda_-}) p_{i,j_-}$$

$$\hat{P}_{i,+} = \sum_{j-}^{n_-} q_{i,j_+} p_{i,j_+} = \sum_{j_+}^{n_+} \frac{1}{Z_+} \exp(\frac{-p_{i,j_+}}{\lambda_+}) p_{i,j_+}$$

最终为了平滑整个曲线，作者加入了hinge loss：

$$\ell_{\text{smooth}}(z) = \frac{1}{L} \log(1 + \exp(Lz))$$

最终分类的loss为：

$$\min_{\theta} \mathcal{L}_{\text{DR}}(\theta) = \sum_{i}^{N} \ell_{\text{smooth}}(\hat{P}_{i,-} - \hat{P}_{i,+} + \gamma)$$

并且为了确保正样本和负样本能够分开，需要

$$\gamma = 0.5$$

即可保证：

$$\forall i, j_+ \quad p_{i,j_+} > 0.5; \quad \forall i, j_- \quad p_{i,j_-} \leq 0.5$$

对于回归loss，作者也做了改进：

$$\ell_{\text{reg}}(x) = \begin{cases} 0.5x^2/\beta & x \leq \beta \\ |x| - 0.5\beta & x \geq \beta \end{cases}$$

主要是在训练中对其进行衰减：

$$\beta_t = \beta_0 - \alpha(t\%K)$$

目的是为了减少L1和L2之间的gap。

三、实验结果

作者首先做了一些消融实验，例如对于hingle loss中的L:

Table 1. Comparison of the smooth term L in Eqn. 8. Training uses 1× iterations and ResNet-101 as the backbone.

L	τ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
5	3	38.6	58.3	41.7	21.5	43.0	51.4
6	5	38.7	58.8	41.5	21.1	42.9	52.1
7	5	38.7	58.9	41.4	21.7	42.9	52.0
8	5	38.6	58.7	41.3	21.6	42.4	51.9

L对最后的结果影响不大。

其他关于正则项h等消融实验详情见论文。

Table 5. Comparison with the state-of-the-art methods on COCO test set.

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
two-stage detectors							
Faster R-CNN+++ [12]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [14]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Deformable R-FCN [3]	Aligned-Inception-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
Mask R-CNN [11]	Resnet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
one-stage detectors							
YOLOv2 [20]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [17]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [6]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [15]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [15]	ResNeXt-101-FPN	40.8	61.1	44.1	24.1	44.2	51.2
$\text{Dr.Retina}_{\text{fixed}}$	ResNet-101-FPN	40.6	60.7	43.9	22.9	43.7	51.9
Dr.Retina	ResNet-101-FPN	41.1	60.7	44.3	23.3	44.1	52.6
Dr.Retina	ResNeXt-101-FPN	**42.5**	**62.8**	**45.9**	**25.2**	[illegible]	[illegible]

最终结果的性能提升还是非常明显的，比baseline高了2个点左右。

四、总结分析

优点：借鉴了排序loss引入到目标检测中，并且给出了可行的优化过程，性能提升也很明显，对于现有的检测框架，只需要修改损失函数，后续会考虑尝试下。

缺点：超参还是蛮多的，虽然在COCO上似乎影响不大，换个数据集和检测框架（例如anchor-free的）不知道是不是很稳定。之前在FCOS上尝试了GHM的方法，直接用默认参数可以掉10个点，不过可以通过调参调回来。