Soft Sampling：探索更有效的采样策略

Anchor-based的目标检测器通过anchor来得到一系列密集的候选框，然后按照一定阈值将候选框分成正样本(前景)和负样本(背景)，最后按照一定的采样策略来进行训练。目标检测中广泛采用的采样策略是随机采样(正样本和负样本按照一定比例随机采样)，然而，随机采样并不能保证能够选取到更有价值的样本(使检测器更鲁棒)。对于目标检测器来说，如何得到更有价值的样本(或者说什么样的样本更有价值)，仍然是一个亟待解决的问题。

在探索更有效的采样策略的过程中，产生了两大类方法：

Hard Sampling：从所有候选样本中选择子集来训练模型。(包括hard negative mining、OHEM、IoU-balanced Sampling等等)

Soft Sampling：为所有候选样本安排不同的权重值。(包括Focal Loss、GHM、PISA等等)

本文主要介绍Soft Sampling的采样策略(包括Focal Loss、GHM和PISA)，且只涉及Soft Sampling部分，其他细节就不具体展开了~~

GHM可以看成是Focal Loss的升级版，而PISA是另外一种Soft Sampling的思路。

1.Focal Loss

这篇文章鼎鼎大名，获得了2017年ICCV最佳学生论文奖，文章中提出的RetinaNet是最近流行的anchor-free检测器的标配，focal loss的提出引发了Soft Sampling研究的热潮，后面对focal loss的分析写的很出彩，附录也值得一读，我觉得GHM可能是受到了Focal Loss附录的启发，总而言之，值得精读几遍~~(一顿吹，哈哈哈)。

Motivation

作者提出前景和背景类别的极度不平衡是造成单阶段检测器精度差的主要原因。

类别的不平衡会产生两个问题：

1.大量容易负样本(不提供有用的学习信息)会导致训练过程无效。

2.大量容易负样本产生的loss会压倒少量正样本的loss(即容易负样本的梯度占主导)，导致模型性能衰退。

作者提出Focal Loss来处理类别不平衡问题。

Cross Entropy

首先定义二分类的cross entropy (CE) loss为：

$$\mathrm{CE}(p, y)=\begin{cases}-\log(p) & \text{if } y=1\\ -\log(1-p) & \text{otherwise}\end{cases}$$

上式的$y\in\{\pm 1\}$特指gt的类别，$p\in[0,1]$是模型估计的标签为$y=1$的概率。为了方便起见，定义p_t为：

$$p_{\mathrm{t}}=\begin{cases}p & \text{if } y=1\\ 1-p & \text{otherwise}\end{cases}$$

并且重写cross entropy (CE) loss为$\mathrm{CE}(p, y)=\mathrm{CE}\left(p_{\mathrm{t}}\right)=-\log\left(p_{\mathrm{t}}\right)$。

预测值和标签相近，loss就低；预测值和标签远离，loss就高。

Balanced Cross Entropy

处理类别不平衡的一般方法为引入权重因子$\alpha \in [0,1]$，类别为1的权重值为α，类别为-1的权重值为$1-\alpha$。为了方便起见，如定义p_t一样的方式定义α_t，α-balanced CE loss为：

$$\mathrm{CE}\left(p_{\mathrm{t}}\right)=-\alpha_{\mathrm{t}} \log \left(p_{\mathrm{t}}\right)$$

Focal Loss Definition

虽然α能够平衡正样本和负样本的重要性，但是不能平衡容易样本和困难样本的重要性。因此，作者修改了loss函数来降低容易样本的权重值，使得训练过程更加关注hard negative examples(即预测值背离标签的样本)。

形式上，在cross entropy loss前面添加一个调制因子$(1-p_{\mathrm{t}})^{\gamma}$，且$\gamma \geq 0$，定义focal loss为：

$$\mathrm{FL}\left(p_{\mathrm{t}}\right)=-(1-p_{\mathrm{t}})^{\gamma} \log \left(p_{\mathrm{t}}\right)$$

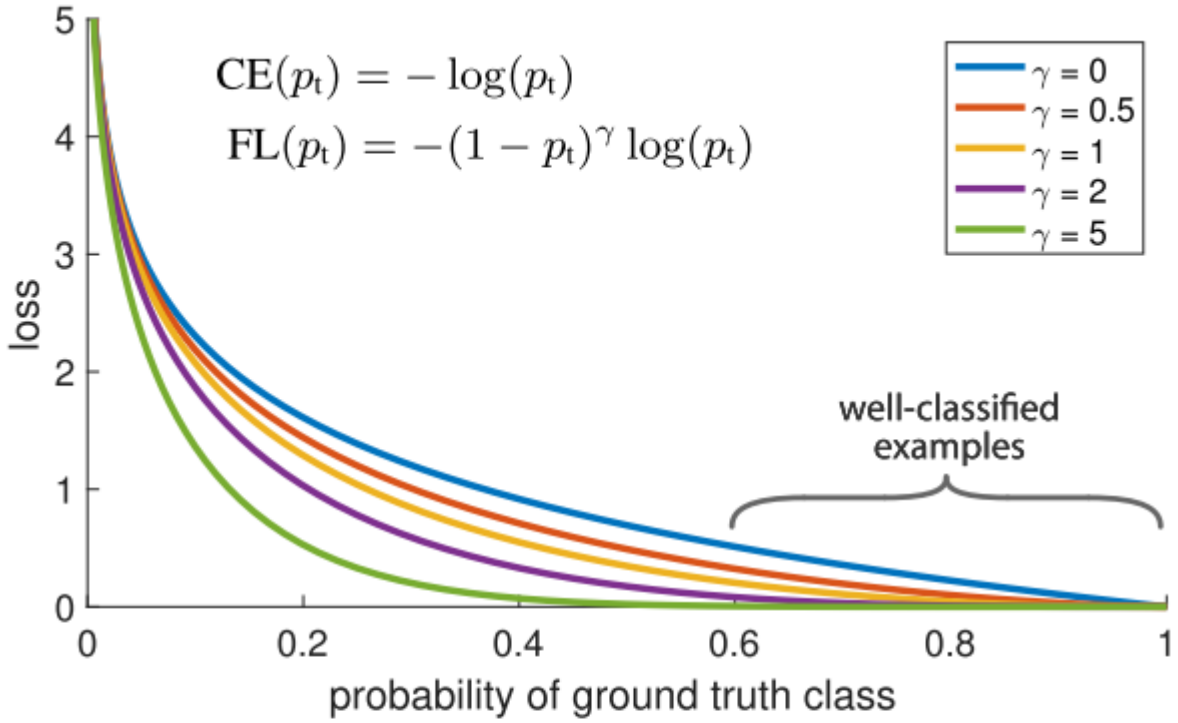


Figure 1. We propose a novel loss we term the *Focal Loss* that adds a factor $(1-p_{\mathrm{t}})^{\gamma}$ to the standard cross entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples ($p_{\mathrm{t}} > .5$), putting more focus on hard, misclassified examples. As our experiments will demonstrate, the proposed focal loss enables training highly accurate dense object detectors in the presence of vast numbers of easy background examples.

focal loss有两个性质：

1.当样本错误分类时，p_{t}偏小，调制因子接近于1并且loss不受影响的。当$p_{\mathrm{t}} \rightarrow 1$时，调制因子趋向于0，对于容易分类样本来说，等同于loss的权重值降低。

2.聚焦参数γ平滑地调整容易样本降低权重值的速率。当$\gamma=0$时，FL等同于CE，随着γ的增加，调制因子的影响会同步的增加。

直观的来看，调制因子能够减少容易样本的loss，并且扩大得到低loss值的样本范围。当调制因子增加时，会增加错误分类样本的重要性。

实际使用时，使用的是α-balanced变体的focal loss：

$$\mathrm{FL}\left(p_{\mathrm{t}}\right)=-\alpha_{\mathrm{t}}\left(1-p_{\mathrm{t}}\right)^{\gamma} \log \left(p_{\mathrm{t}}\right)$$

作者还从实验中发现，focal loss的形式其实不重要。在附录中，作者还考虑了focal loss的其他形式，并且证明其他形式同样有效。

focal loss在实际使用中还有两个重要的小细节：

1.一张图片总的focal loss是对~100K的anchors求focal loss的和得到的，然后通过一个gt box的anchors数量进行归一化。因为大量的anchors是容易负样本，在focal loss下得到的loss值非常小，如果用所有anchors的数量进行归一化，会导致归一化后的loss值非常小。

)2.当只使用α时，将α偏向于样本少的类别；当同时使用α和γ时，α和γ需要向相反的方向变化(实验中最佳的参数设置为$\alpha=0.25$和$\gamma=2$)。

Analysis of the Focal Loss

为了更好的理解focal loss，作者仔细分析了loss的经验分布。采用默认的ResNet101，600输入，$\gamma=2$的模型。作者将这个模型使用到大量的随机图片上，并且采样得到$\sim 10^{7}$个负样本和$\sim 10^{5}$个正样本的预测概率。然后，将正负样本分开，计算出这些样本的FL并且归一化loss使得和为一。得到归一化的loss后，就能够从小到大排序loss并且绘制出cumulative distribution function (CDF)。

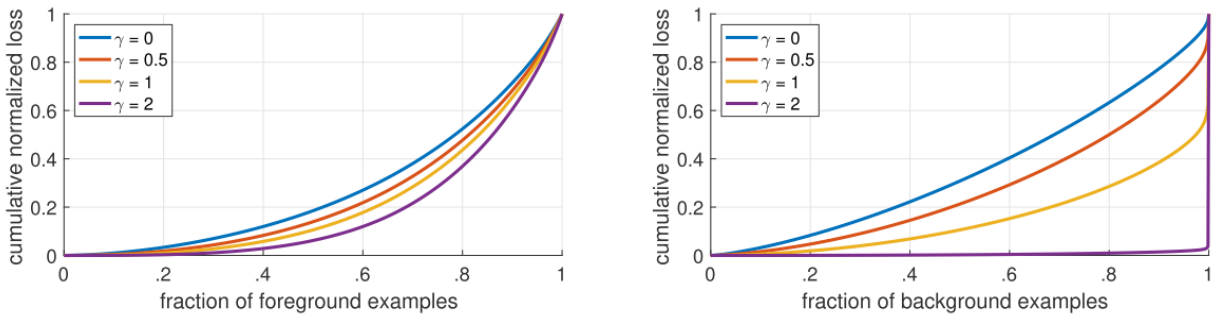


Figure 4. Cumulative distribution functions of the normalized loss for positive and negative samples for different values of γ for a *converged* model. The effect of changing γ on the distribution of the loss for positive examples is minor. For negatives, however, increasing γ heavily concentrates the loss on hard examples, focusing nearly all attention away from easy negatives.

正样本在不同γ下的CDF基本相同。举个例子，大约有20%的最难正样本占据了大约一半的正样本loss，随着γ的增加，更多的loss集中在20%的最难正样本上，但是影响很小。

而负样本在不同γ下的CDF表现完全不同。当$\gamma=0$时，正样本和负样本的CDFs非常相似。然而，随着γ的增加，更多的权值集中在困难样本上。事实上，$\gamma=2$时，大量的loss来自于小部分的负样本。

从以上分析可以看出，FL可以有效地降低容易样本的影响，将所有的注意力集中在困难样本上。

Focal Loss*

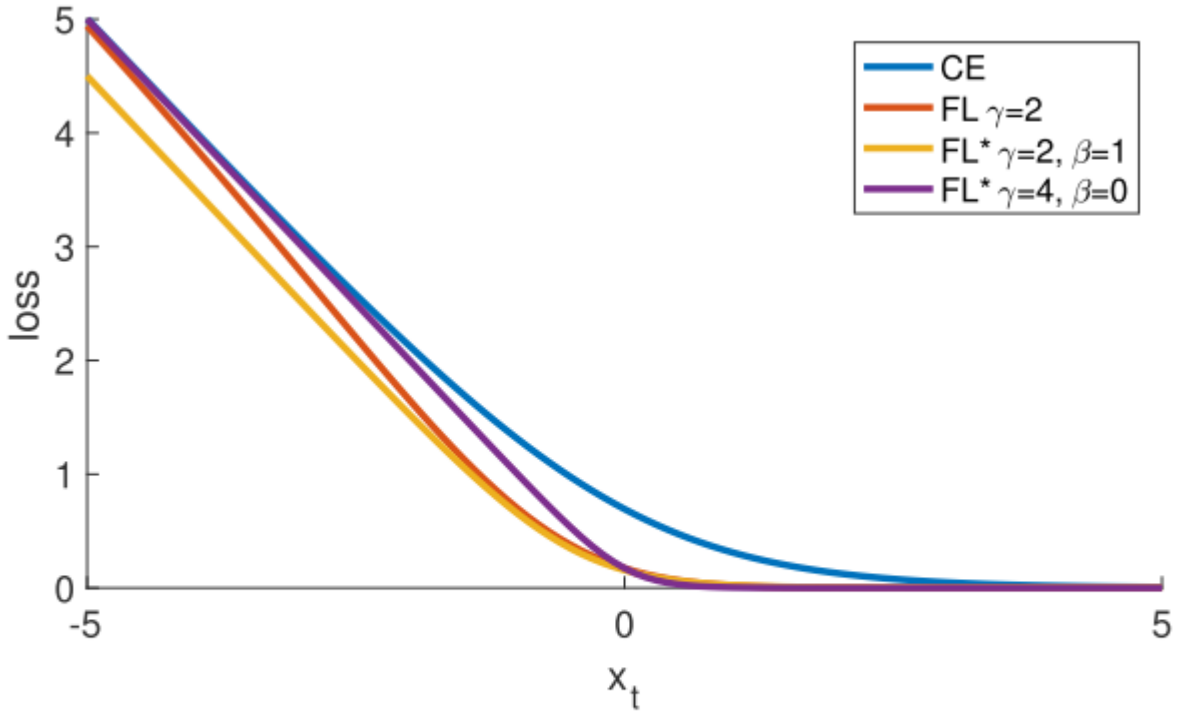


Figure 5. Focal loss variants compared to the cross entropy as a function of $x_{\mathrm{t}} = yx$. Both the original FL and alternate variant FL^* reduce the relative loss for well-classified examples ($x_{\mathrm{t}} > 0$).

focal loss具体形式并不是关键。作者提出了另外一种focal loss的形式能够取得可比较的结果。

作者首先考虑与上文有所不同的CE和FL的形式。定义数值x_{t}如下：

$$x_{\mathrm{t}} = yx$$

$y \in \{\pm 1\}$特指gt类别。定义$p_{\mathrm{t}} = \sigma\left(x_{\mathrm{t}}\right)$($p_{\mathrm{t}}$与之前定义的$p_{\mathrm{t}}$是一致的)。当$x_{\mathrm{t}} > 0$时，样本被正确的分类，这种情况下$p_{\mathrm{t}} > .5$。

使用x_t就能够定义一个focal loss的替换形式。p_{t}^*和FL^*如下：

$$\begin{aligned} p_{\mathrm{t}}^* &= \sigma\left(\gamma x_{\mathrm{t}} + \beta\right) \\ \mathrm{FL}^* &= -\log(p_{\mathrm{t}}^*)/\gamma \end{aligned}$$

FL^*有两个参数γ和β，能够用来控制loss曲线的陡度(steepness)和位移(shift)。和FL一样，FL^*能够降低容易样本的权值。采用和上述FL一样的设置，模型能够实现相当的精度。

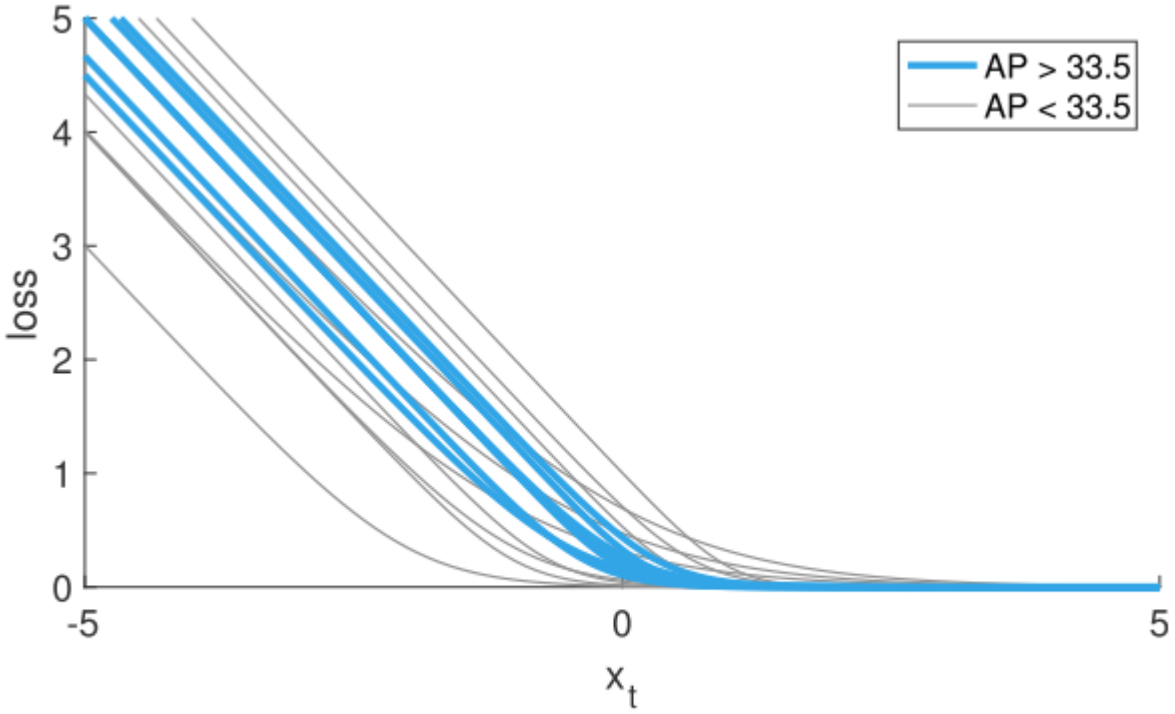


Figure 7. Effectiveness of FL* with various settings γ and β. The plots are color coded such that effective settings are shown in blue.

蓝色的线表示AP大于33.5的有效设置。如图所示，容易样本($x_t > 0$)的loss下降。更一般的，任何和FL 、FL^*有相似性质的函数同样有效。

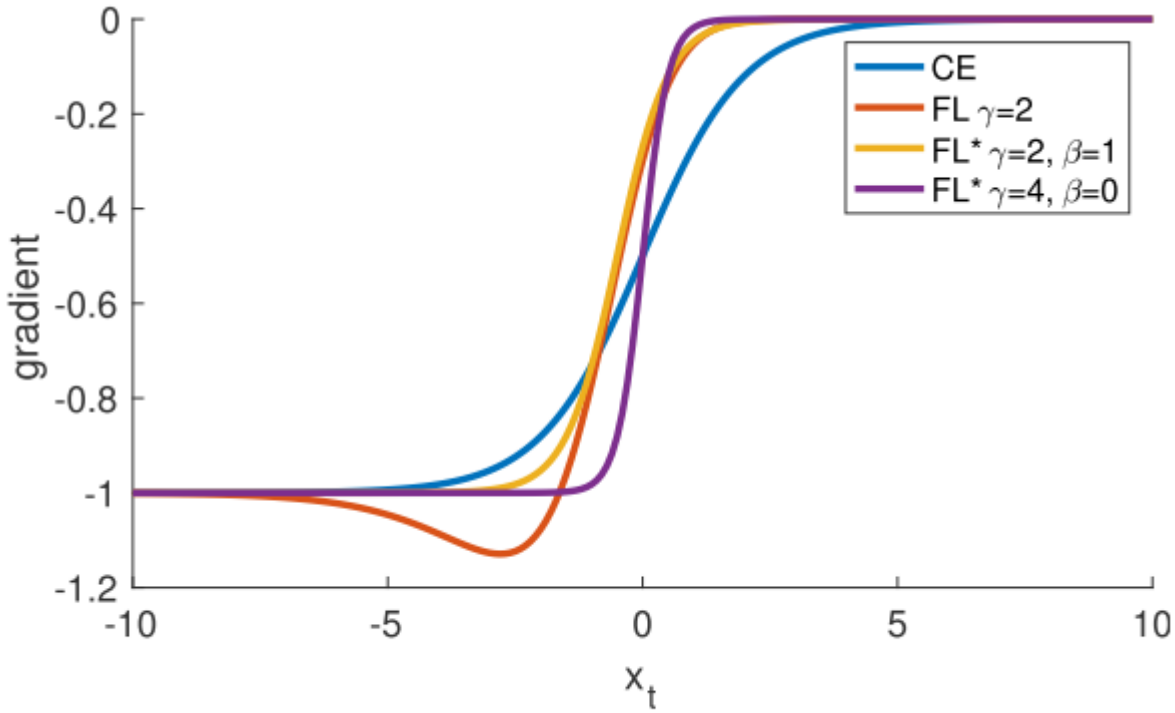


Figure 6. Derivates of the loss functions from Figure 5 w.r.t. x.

$\mathrm{CE}, \mathrm{FL},$和 FL^*的导数为：

$$\frac{d\mathrm{CE}}{dx} = y\left(p_t - 1\right) \quad \frac{d\mathrm{FL}}{dx} = y(1-p_t)^{\gamma}\left(\gamma p_t \log(p_t) + p_t - 1\right) \quad \frac{d\mathrm{FL}^*}{dx} = y\left(p_t^* - 1\right)$$

所有loss函数，随着预测置信度的增加梯度趋向于-1或者0。然而，当$x_t > 0$时，FL 和 FL^*梯度非常小。

2.GHM

由Focal Loss附录的图可以看出，当梯度绝对值越大时，loss值也越大(即梯度和困难样本存在某种关系)，这一点可能启发了GHM的工作。为了降低复杂度，使用了unit region approximation，该部分不具体展开了。

Motivation

目标检测器中存在着两个问题：正负样本的不平衡和难易样本的不平衡。

Focal Loss虽然部分解决了这两个问题，但是Focal Loss采用了2个超参数，需要大量实验进行调参。

如果从整体来看，大量负样本通常是容易样本，大量正样本通常是困难样本，因此，两种不平衡可以粗略的归结为属性不平衡(或者说类别不平衡)。

作者指出类别的不平衡可以归结为难度的不平衡，难度的不平衡可以归结为梯度范数分布的不平衡。

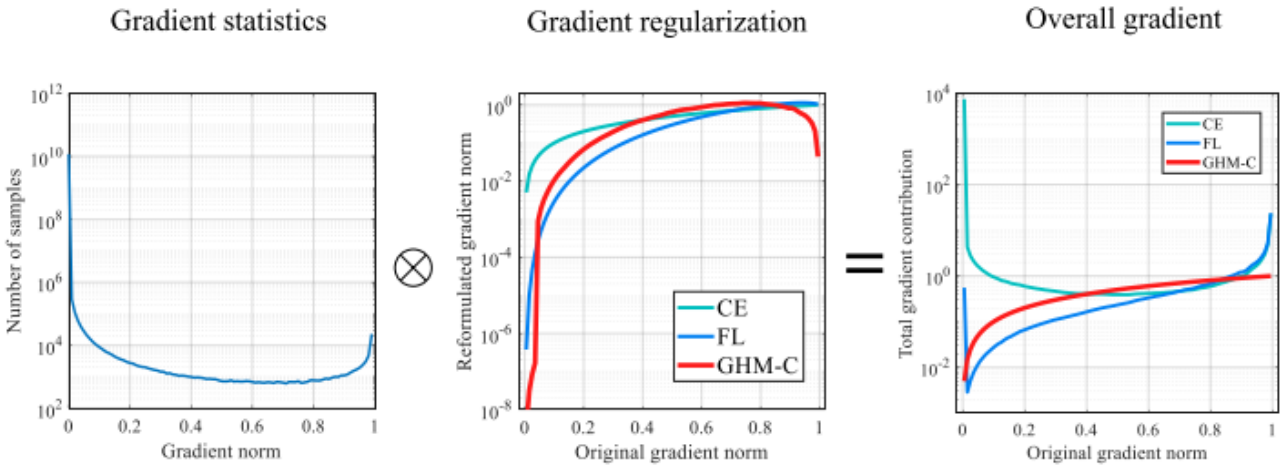


Figure 1: An illustration of gradient harmonizing mechanism. The figure in the left displays the distribution of relative gradient norm in a converged model in log scale respectively. The middle figure displays the new gradient norms after the rectification of Focal Loss (FL) and GHM-C loss, compared with the original cross-entropy (CE) loss. The right figure shows the total gradient contribution of examples w.r.t gradient norm.

可以通过梯度范数分布来表示不同属性的不平衡。为了方便起见，将关于梯度范数的样本密度称为梯度密度。梯度范数非常小的样本占据着相当大的密度。虽然容易样本对于全局梯度的贡献少于困难样本，但是大量容易样本总的贡献压倒了少量困难样本总的贡献。另外，作者还发现梯度范数非常大的样本的密度略大于介于中间的样本的密度。作者认为这些**非常困难的样本可以当作异常样本**，因为当模型稳定时，这些异常样本仍然存在。这些异常样本可能影响模型的稳定性，因为异常样本的梯度可能和其他样本的梯度有巨大差异。如果收敛模型被迫去学习如何更好地对这些异常样本进行分类，那么会导致其他样本分类精度的衰退。

通过上述分析，作者提出一个梯度协调机制(gradient harmonizing mechanism简写为GHM)来有效训练检测器，该**梯度协调机制可以在训练过程中调整不同样本的梯度贡献。GHM首先对梯度密度相似的样本的数量进行统计，然后根据梯度密度在各样本的梯度上附加一个协调参数。**利用GHM进行训练时，容易样本产生的大量累积梯度可以很大程度上降低权值，另外异常值也可以相对降低权值。最终导致每种样本的贡献会趋于平衡，训练过程会更加有效和稳定。

实际中，梯度的修改可以通过重新设计损失函数来等效地实现，我们将GHM嵌入到分类损失中，该分类损失被称为GHM-C loss。 这种损失函数不需要超参数进行调整。 **由于梯度密度是一个统计变量，取决于小批量的样本分布，所以GHM-C是一种动态loss，可以适应每个批次的数据分布变化以及模型的更新。** 为了显示出GHM的通用性，在回归分支中利用它来构成GHM-R loss。

Cross Entropy

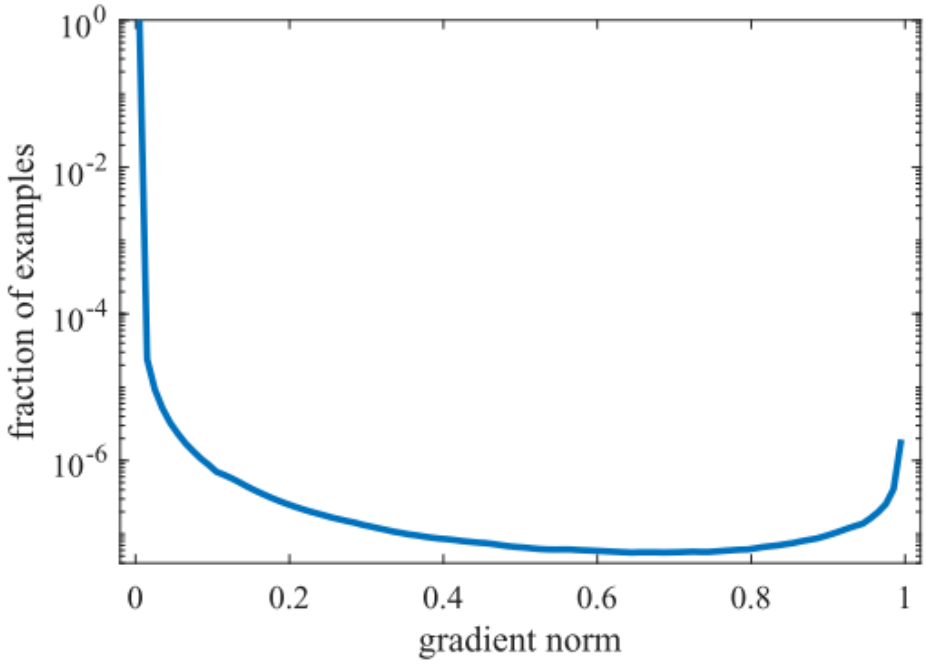


Figure 2: The distribution of the gradient norm g from a converged one-stage detection model. Note that the y-axis uses log scale since the number of examples with different gradient norm can differ by orders of magnitude.

首先定义二分类的cross entropy (CE) loss为：

$$L_{CE}\left(p,p^*\right)=\begin{cases}-\log(p) & \text{if } p^*=1\\ -\log(1-p) & \text{if } p^*=0\end{cases}$$

其中$p\in[0,1]$是模型预测出来的概率，而$p^*\in\{0,1\}$为gt的类标签。

设x为模型的直接输出，于是可将p定义为$p=sigmoid(x)$，对x求梯度：

$$\frac{\partial L_{CE}}{\partial x}=\begin{cases}p-1 & \text{if } p^*=1\\ p & \text{if } p^*=0\end{cases}$$
$$=p-p^*$$

g定义如下：

$$g = |p - p^*| = \begin{cases} 1-p & \text{if } p^* = 1 \\ p & \text{if } p^* = 0 \end{cases}$$

g等于x的梯度范数。g值表示样本的属性并且隐含着样本对全局梯度的影响。

Gradient Density

梯度密度函数定义如下：

$$GD(g) = \frac{1}{l_\epsilon(g)} \sum_{k=1}^{N} \delta_\epsilon(g_k, g)$$

其中g_k是第k个样本的梯度范数，那么$\delta_\epsilon(x, y)$和$l_\epsilon(g)$定义为：

$$\delta_\epsilon(x, y) = \begin{cases} 1 & \text{if } y - \frac{\epsilon}{2} <= x < y + \frac{\epsilon}{2} \\ 0 \end{cases} \qquad l_\epsilon(g) = \min\left(g + \frac{\epsilon}{2}, 1\right) - \max\left(g - \frac{\epsilon}{2}, 0\right)$$

g的梯度密度是以g为中心ϵ为宽范围内的样本数量通过有效范围的宽度进行归一化得到的。

接着定义梯度密度协调参数为：

$$\beta_i = \frac{N}{GD(g_i)}$$

N是样本总数。为了更好的理解梯度密度协调参数，可以将β_i改写成$\beta_i = \frac{1}{GD(g_i)/N}$。分母$GD(g_i)/N$是对梯度位于$g_i$范围的部分样本进行归一化。如果所有样本的梯度是统一分布的，那么对于任意g_i都有$GD(g_i) = N$，并且每个样本都有相同的$\beta_i = 1$，这意味着没有任何改变。如果样本具有大的梯度密度，那么归一化后会导致权值下降。

GHM-C Loss

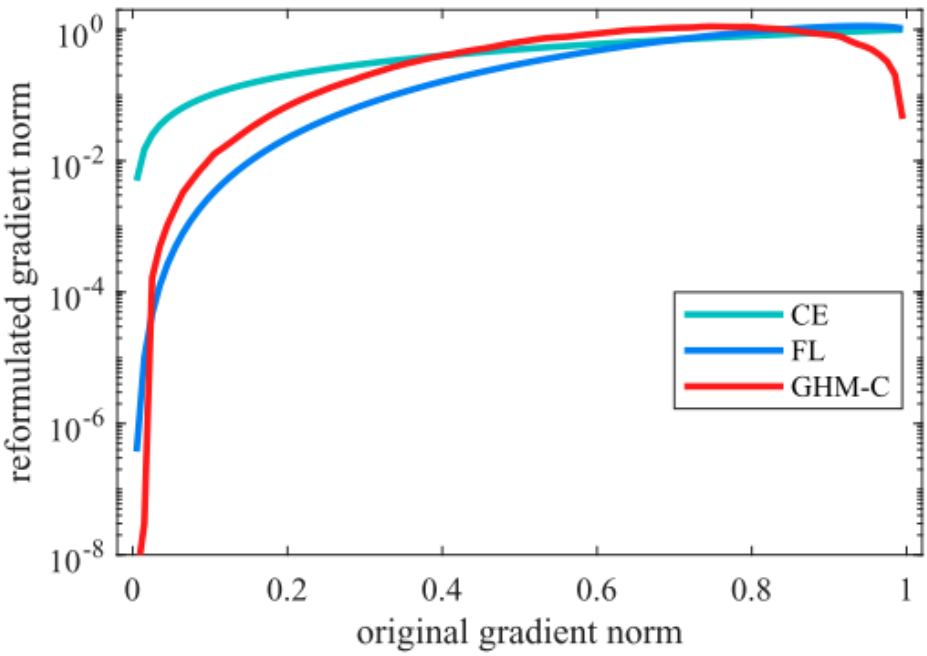


Figure 3: Reformulated gradient norm of different loss functions w.r.t the original gradient norm g. The y-axis uses log scale to better display the details of FL and GHM-C.

通过将β_i作为loss的权值嵌入到分类loss中，梯度密度协调的分类loss定义如下：

$$L_{GHM-C} = \frac{1}{N}\sum_{i=1}^{N}\beta_i L_{CE}\left(p_i, p_i^*\right) = \sum_{i=1}^{N}\frac{L_{CE}\left(p_i, p_i^*\right)}{GD\left(g_i\right)}$$

从图中可以看到，Focal Loss和GHM-C有相似的趋势，这说明超参数最优的Focal Loss与梯度均匀协调的GHM曲线相似。另外，GHM-C比起Focal Loss多了一个优点：降低异常样本的权值。

GHM-R Loss

将回归分支预测得到的参数化偏移量定义为$t = (t_x, t_y, t_w, t_h)$并且将gt的目标偏移量定义为$t^* = \left(t_x^*, t_y^*, t_w^*, t_h^*\right)$。回归loss采用smooth L1 loss:

$L_{reg} = \sum_{i\in\{x,y,w,h\}} SL_1\left(t_i - t_i^*\right)$

其中$SL_1(d) = \begin{cases} \frac{d^2}{2\delta} & \text{if } |d| <= \delta \\ |d| - \frac{\delta}{2} & \text{otherwise} \end{cases}$

δ是二次曲线部分和线性部分的分割点，实际中，通常设置为1/9。

令$d = t_i - t_i^*$，那么smooth L1 loss关于t_i的梯度为：

$\frac{\partial SL_1}{\partial t_i} = \frac{\partial SL_1}{\partial d} = \begin{cases} \frac{d}{\delta} & \text{if } |d| <= \delta \\ \text{sgn}(d) & \text{otherwise} \end{cases}$

所有$|d|$超过分割点的样本都有相同的梯度范数$\left|\frac{\partial SL_1}{\partial t_i}\right| = 1$，这使得依赖梯度范数来区分具有不同属性的样本是不可能的。一种替代方法是直接使用$|d|$作为不同属性的衡量标准，但是新的问题是$|d|$理论上能够无限大，导致unit region approximation无法实现。

为了方便的将GHM用到回归loss上，作者将SL1 loss修改为更加优雅的形式：

$ASL_1(d) = \sqrt{d^2 + \mu^2} - \mu$

ASL1 loss和SL1 loss共享相同的属性：当d的绝对值比较小时，近似为L2 loss，当d的绝对值比较大时，近似为L1 loss。ASL1 loss处处可导连续，而SL1 loss在$d = \delta$处不可导。ASL1 loss关于d的梯度为：

$\frac{\partial ASL_1}{\partial d} = \frac{d}{\sqrt{d^2+\mu^2}}$

梯度范围为$[0, 1)$，因此计算ASL1 loss的梯度密度和CE loss一样方便。

定义$gr = \left|\frac{d}{\sqrt{d^2+\mu^2}}\right|$作为ASL1 loss的梯度范数。

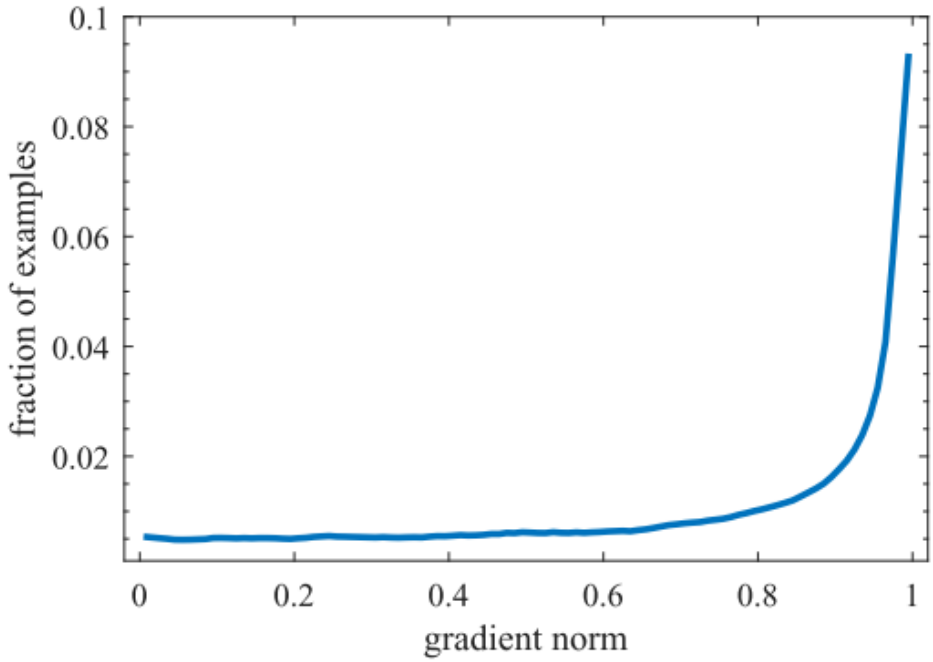


Figure 4: The distribution of the gradient norm gr for ASL_1 loss.

从图中可以看出，存在大量的异常样本，因为回归只对正样本进行，所以分类和回归之间的差异趋势是合理的。

$$
\begin{aligned}
L_{GHM-R} &= \frac{1}{N}\sum_{i=1}^{N}\beta_i ASL_1\left(d_i\right) \\
&= \sum_{i=1}^{N}\frac{ASL_1\left(d_i\right)}{GD\left(gr_i\right)}
\end{aligned}
$$

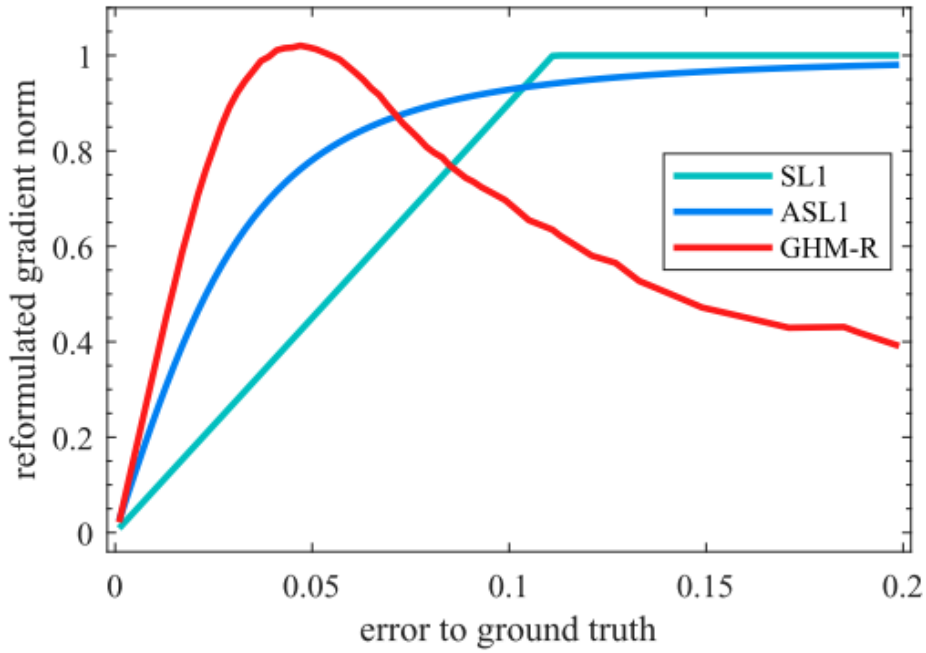


Figure 5: Comparison of the reformulated gradient contributions of different regression losses w.r.t the value of $|d|$, i.e. the error to ground-truth.

作者认为，在回归中，并不是所有"容易样本"都不重要。 分类中的容易样本通常是具有非常低的预测概率的背景区域，并且肯定会被排除在最终的候选之外。 因此，这种样本的改善对精度几乎没有影响。 但是在回归中，一个容易样本仍然偏离了gt的位置。 更好地预测任意样本会改善最终候选框的质量。 此外，先进的数据集更加关注定位的精度。通过提高容易样本的权值和降低异常样本的权值，GHM-R loss能够调整容易样本和困难样本对回归的影响。

3.PISA

PISA跳出了Focal Loss的思路，认为应当从mAP这个指标出发，来设计采样策略。该论文分析部分丝丝入扣，值得仔细推敲。

Motivation

虽然随机采样和hard sampling简单且广泛采用，但不一定是训练有效探测器的最佳采样策略。

一个开放的问题是：用来训练目标探测器的最重要的样本是什么？作者揭示了设计采样策略时需要考虑两个重要方向：

1.**样本不应该被认为是独立的和同等重要的。**基于区域的目标检测是从大量候选框中选取一小部分边界框，以覆盖图像中的所有目标。因此，不同样本的选择是相互竞争的，而不是独立的。一般来说，检测器更可取的做法是在确保所有感兴趣的目标都被充分覆盖时，在每个目标周围的边界框产生高分，而不是对所有正样本产生高分。作者研究表明关注那些与gt目标有最高IOU的样本是实现这一目标的有效方法。

2.**目标的分类和定位是有联系的。**准确定位目标周围的样本非常重要，这一观察具有深刻的意义，即目标的分类和定位密切相关。具体地，定位好的样本需要具有高置信度好的分类。

受上述研究的启发，作者提出了PrIme Sample Attention(PISA)，这是一种简单而有效的方法，用于对区域进行采样和学习目标检测器，作者将那些起着更重要作用的样本作为主要样本。 具体而言，我们定义IoU分层局部排序(IoU Hierarchical Local Rank简写为IoU-HLR)来对每个小批量中的样本进行排序。 该排序策略将每个目标周围的最高IoU的样本放置在排序列表的顶部，并通过简单的权值重标定方案将训练过程的重点引导到排序列表的顶部。作者还设计了分类感知的回归损失来共同优化分类和回归分支。 具体地，这个loss将抑制掉那些回归loss大的样本，从而加强了对主要样本的注意。

Prime Samples

引入Prime Samples的概念，指那些对目标检测性能有重大影响的样本。具体地，作者通过回顾mAP指标来研究不同样本的重要性，研究显示每个样本的重要性取决于它和gt的IoU。因此，作者提出一种新颖的IoU-HLR排序策略，作为一种评估重要性的定量方式。

A Revisit to mAP

给定一个带标注gt的图片，当边界框满足以下两个条件时，标记为正样本。

1.边界框和最近的gt的IoU大于阈值θ时

2.如果gt没有IoU大于阈值的边界框，则选择IoU最大的边界框

剩余没有标记的边界框作为负样本。

mAP指标揭露了对于目标检测器来说更重要样本的**2个准则**:

1.在所有和gt目标重合的边界框中，IoU最高的边界框时最重要的，因为它的IoU值直接影响召回率。

2.所有不同目标的最高IoU边界框中,具有更高的IoU的边界框更加重要,因为它是随着θ增加最后一个低于阈值θ的边界框,从而对整体精度有很大的影响。

IoU Hierarchical Local Rank (IoU-HLR)

基于上述分析，作者提出了IoU-HLR来对小批量边界框样本的重要性进行排序。这个排序是以层次的方式计算的，它既反映了局部的IoU关系(每个ground truth目标周围)，也反映了全局的IoU关系(覆盖整个图像或小批图像)。值得注意的是，不同于回归前的边界框坐标，IoU-HLR是根据样本的最终定位位置来计算的，因为mAP是根据回归后的样本位置来计算的。

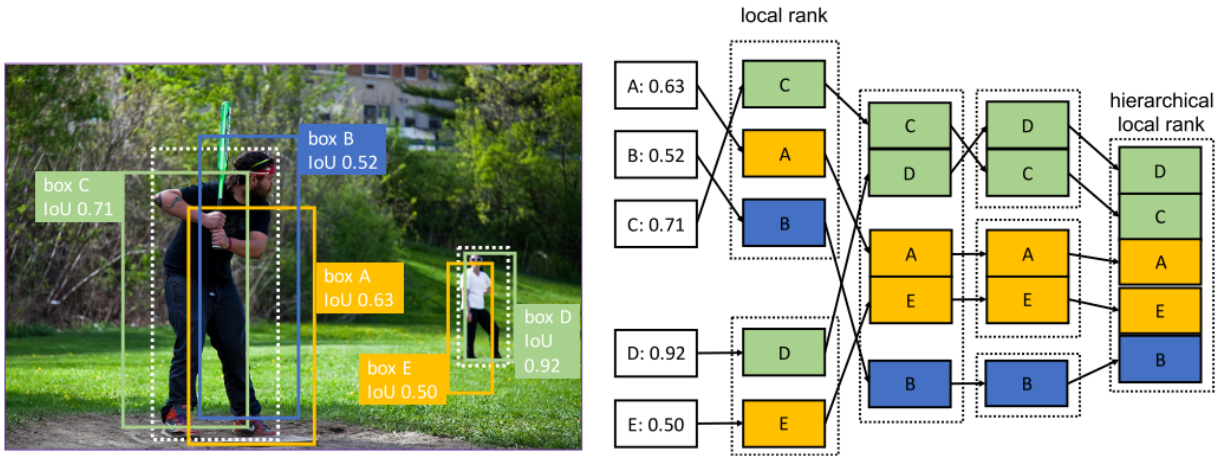


Figure 3: Two steps to compute IoU-HLR. Samples are first sorted by IoU locally, and then sorted in the same-rank group.

我们首先将所有的样本根据它们最近的gt目标分成不同的组。接下来，我们根据每个组的IoU和gt按降序排列样本，得到IoU-LR。随后，我们使用相同的IoU-LR采样，并按降序排序。具体来说，所有top1的IoU-LR样本都被收集和排序，然后是top2、top3等等。这两个排序阶段导致了一个批次的所有样本之间的线性顺序，即IoU-HLR。

IoU-HLR遵循上述两个准则。首先，它将那些局部排序较高的样本放在前面，这些样本对于每一个单独的gt目标来说是最重要的。其次，在每个局部组内，根据IoU对样本进行重新排序，这符合第二个准则。值得注意的是，那些排在列表前面的样本，通常能够确保高精度，因为它们直接影响召回率和精确度，尤其是当IoU阈值较高时;在实现高检测性能时，列表后面的样本就不重要了。

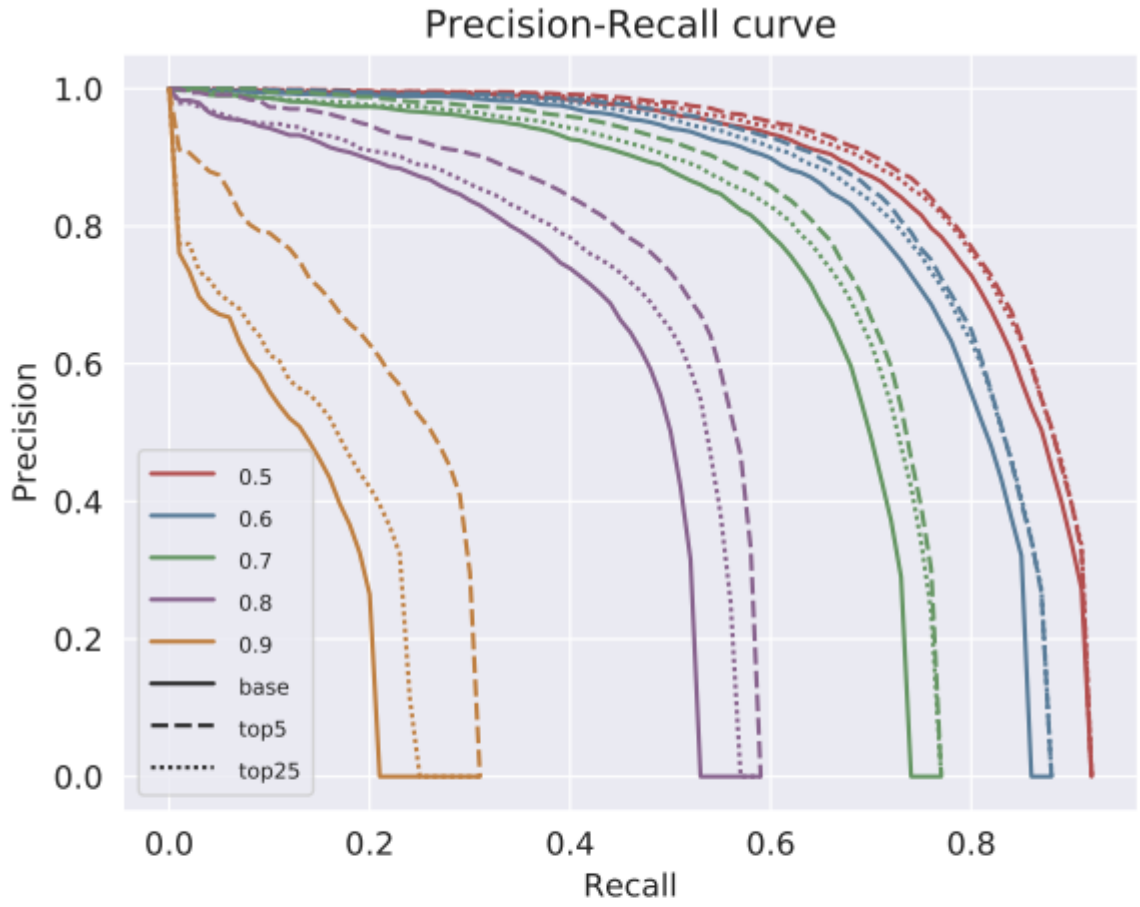


Figure 2: Precision-recall curve under different IoU thresholds. The solid lines correspond to the baseline, dashed lines and dotted lines are results of reducing the classification loss by increasing scores of positive samples. Top5 and top25 IoU-HLR samples are concentrated on respectively.

在相同的情况下，例如，总损失减少10%，我们增加IoU-HLR下top5和top25的样本的得分，并分别用虚线和虚线绘制结果。结果表明，只关注最重要的样本比平等关注更多的样本要好。

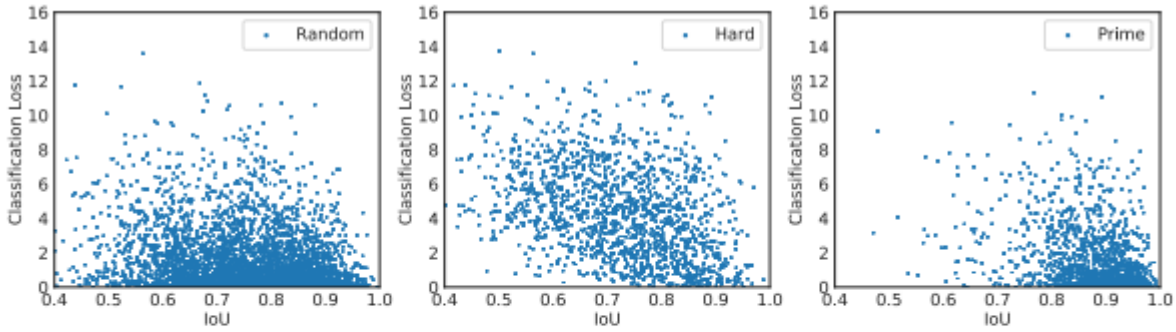


Figure 4: The distribution of random, hard, and prime samples. Here, the hard samples are chosen as the ones with top three loss values from each image; while the prime samples are those ranked as top three according to IoU-HLR.

图中绘制出了random、hard、prime samples的分布。可以观察到，prime samples往往同时有高IoU和低分类loss，而hard samples往往有更高的分类loss并沿IoU轴分布范围更广。表明这两个采样策略具有本质上不同的特性。

Learn Detectors via Prime Sample Attention

作者提出Prime Sample Attention，一种简单且有效的采样策略，该采样策略将更多的注意力集中到主要样本上。PISA由两个组件组成：基于重要性的样本权值重标定(Importance- based Sample Reweighting简写为ISR)和分类感知的回归损失(Classification Aware Regression Loss简写为CARL)。PISA在训练过程中更关注主要样本，而不是平等对待所有样本。首先，主要样本的loss权值大于其他样本，因此分类器倾向于对这些样本预测更高的分数。其次，构造一个联合目标对分类和回归进行学习，从而提高主要样本的分数。

Importance-based Sample Reweighting

IoU-HLR作为重要性的衡量指标，首先将排序转化为线性映射的真实值。根据线性映射的定义，IoU-HLR在每个类中分别进行计算。对于类j，假设总共有n_j个样本，通过IoU-HLR表示为$\{r_1, r_2, \ldots, r_{n_j}\}$，其中$0 \leq r_i \leq n_j - 1$，然后使用一个线性函数转换$r_i$为$u_i$，$u_i$表示为类j的第i个样本的重要程度。

$$u_i = \frac{n_j - r_i}{n_j}$$

采用指数的形式来进一步将样本重要性u_i转换为loss权值w_i。γ表明对重要样本给予多大的优先权的程度因子，β是决定最小样本权值的偏差。

$$w_i = ((1-\beta)u_i + \beta)^{\gamma}$$

通过提出的权值重标定策略，可以将交叉熵重写成：

$$L_{cls} = \sum_{i=1}^{n} w_i' CE(s_i, \hat{s}_i) + \sum_{i=n+1}^{m} CE(s_i, \hat{s}_i)$$

$$w_i' = w_i \frac{\sum_{i=1}^{n} CE(s_i, \hat{s}_i)}{\sum_{i=1}^{n} w_i CE(s_i, \hat{s}_i)}$$

n和m是正样本和所有样本的总数，s_i和$\hat{s}_i$表示第i个样本的预测分数和分类目标。值得注意的是，简单的添加loss权值将会改变总的loss值并且改变正样本和负样本的比例，因此，为了保持正样本总的loss值不改变，作者将w_i归一化为w_i'。

Classification-Aware Regression Loss

受到分类和定位是相关的，由此受到启发，作者提出了另一种方法来关注主要样本。作者提出使用分类感知的回归损失(CARL)来联合优化分类和回归两个分支。CARL可以提升主要样本的分数，同时抑制其他样本的分数。回归质量决定了样本的重要性，我们期望分类器对重要样本输出更高的分数。两个分支的优化应该是相互关联的，而不是相互独立的。

作者的解决方法是让回归损失能够感知到分类分数，这样梯度就可以从回归分支传递到分类分支。公式如下：

$$
\begin{aligned}
L_{reg} &= \sum_{i=1}^{n} c_i' \mathcal{L}\left(d_i, \hat{d}_i\right) \\
c_i' &= c_i \frac{\sum_{i=1}^{n} \mathcal{L}\left(d_i, \hat{d}_i\right)}{\sum_{i=1}^{n} c_i \mathcal{L}\left(d_i, \hat{d}_i\right)} \\
c_i &= \frac{v_i}{\frac{1}{n}\sum_{i=1}^{n} v_i} \\
v_i &= ((1-b)p_i + b)^k
\end{aligned}
$$

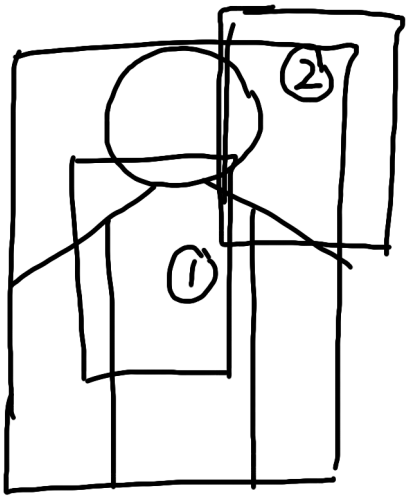
p_i表示相应类别的预测分数，d_i表示输出的回归偏移量。利用一个指数函数将p_i转化为v_i，随后根据所有样本的平均值对它进行缩放。为了保持损失规模不变，对具有分类感知的c_i进行归一化。$\mathcal{L}$是常用的smooth L1 loss。

L_{reg}关于c_i'的梯度与原回归损失$\mathcal{L}\left(d_i, \hat{d}_i\right)$成正比。$L_{reg}$关于$p_i$的梯度与$\mathcal{L}\left(d_i, \hat{d}_i\right)$正相关。即回归损失越大的样本分类得分的梯度越大，说明对分类得分的抑制作用越强。从另一个角度看，$\mathcal{L}\left(d_i, \hat{d}_i\right)$反映了样本i的定位质量，因此可以认为是一个IoU的估计，进一步可以看作是一个IoU-HLR的估计。可以近似认为，排序靠前的样本有较低的回归损失，于是分类得分的梯度较小。对于CARL来说，分类分支受到回归损失的监督。 不重要样本的得分被极大的抑制掉，而对重要样本的关注得到加强。

总结

Focal Loss认为正负的不平衡，本质上是因为难易样本的不平衡，于是通过修改交叉熵，使得训练过程更加关注那些困难样本，而GHM在Focal Loss的基础上继续研究，发现难易样本的不平衡本质上是因为梯度范数分布的不平衡，和Focal Loss的最大区别是GHM认为最困难的那些样本应当认为是异常样本，让检测器强行去拟合异常样本对训练过程是没有帮助的。PISA则是跳出了Focal Loss的思路，认为采样策略应当从mAP这个指标出发，通过IoU Hierarchical Local Rank (IoU-HLR)，对样本进行排序并权值重标定，从而使得recall和precision都能够提升。

我认为PISA的方法有一个问题，就是只考虑了IoU的排序，但是没有考虑到即使是相同IoU的样本的重要程度也是不相同的，比如有两个相同的IoU样本，一个位于目标的内部，另一个没有在目标的内部，显然第一个样本应当比第二个样本重要，如果后面要对PISA方法进行改进的话，可能需要引入一个**有效主要样本**的概念。



Reference

1.https://arxiv.org/abs/1708.02002

2.https://arxiv.org/abs/1811.05181

3.https://arxiv.org/abs/1904.04821

欢迎交流指正~