# 通过ATSS自适应训练样本选择缩小anchor-based和anchor-free差距



# 1. 摘要与介绍

## 1.1. 摘要

anchor-based和anchor-free的主要差别在于**正负样本的划分。**如果anchor-free和anchor-based都使用同一套正负样本划分标准，那么无论是point还是box回归没有什么差距。

于是提出ATSS自适应训练样本选择（Adaptive Training Sample Selection），**根据物体的统计特征自动选择正负样本。**

并在最后讨论了在图像上每个位置**铺设多个锚点来检测物体的必要性。**

## 1.2. 关于anchor-free检测器

anchor-free检测器以两种方式直接找到没有预设anchor box的物体：

- **keypoint-based**：首先定位几个预设或者自我学习到的关键point，然后约束物体的空间范围（CornerNet）。
- **center-based**：利用物体的中心点或区域来定义positive目标，然后预测positive目标到物体边界的四周距离（FCOS）。center-based更类似于anchor-based，它将点作为预设样本而不是anchor box。

anchor-free检测器的好处：能够消除与anchor box有关的超参数，并能够取得与anchor-based检测器相当的性能，使其在泛化能力方面更具潜力。

## 1.3. anchor-based和anchor-free区别

以one-stage anchor-based检测器RetinaNet和center-based anchor-free检测器FCOS为例，他们有**三个区别**：

| | **RetinaNet（one-stage/anchor-based)** | **FCOS（center-based/anchor-free)** |
|---|---|---|
| anchor的数量 | 每个位置铺设几个anchor box | 每个位置铺设一个anchor point |
| 对正负样本的划分 | IoU值划分正负样本 | 利用spatial与scale的限制 *（FPN） |
| 回归的起始状态 | 从预设的anchor box回归bounding box | 从anchor point回归bounding box |

> *** 关于"利用spatial与scale的限制"**
>
> > **FCOS首先用spatial约束来寻找空间维度的候选正样本，然后使用scale约束来选择尺度维度的最终正样本。**
>
> spatial约束是指从所有特征图point约束到所有落入ground-truth box范围内的anchor point。
>
> scale约束是指FPN生成的每层特征图约束了每层检测物体尺寸的范围。具体来说，对于一层特征图，目标的正样本的尺寸必须满足在该层划定的检测尺寸，除此之外均为负样本。
>
> 详情请见FCOS论文。

经过复现实验证明，如果将RetinaNet和FCOS的正负样本划分控制一致，在性能上并没有多大差距。

由此，本文提出一个新的自适应训练样本选择（ATSS），根据物体的特征自动选择正负样本。

## 1.4. 本工作主要贡献

- 表明anchor-free和anchor-based检测器之间的本质区别实际上是如何定义正负训练样本。
- 提出一种自适应训练样本选择，根据对象的统计特征自动选择正负样本训练。
- 证明在图像上每个位置堆砌多个anchor来检测物体时一种无用的操作。
- 获得MS COCO的SOTA。

# 2. anchor-based和anchor-free检测器的差别分析

采用anchor-based的RetinaNet和anchor-free的FCOS来分析差异。

在第二节中，重点讨论对**正负样本的划分**和**回归的起始状态**。

在第三节中，将讨论**anchor的数量**对于性能的影响。

## 2.1. 实验设置

**数据集**

使用包含80个类别的MS COCO。

**训练细节**

- backbone：使用ImageNet预训练的ResNet50的5层特征金字塔结构。对于RetinaNet，每层特征金字塔都有一个尺寸为$8S$的正方形anchor相关联，$S$为总共的步长stride大小。
- 图像大小：短边为800，长边为1333
- 使用随机梯度下降SGD进行了90k迭代
- 动量：0.9
- 权重衰减：0.0001
- batch size：16
- 学习率：初始为0.01，在60k和80k次迭代时分别衰减0.1

**推理细节**

图像大小与训练一致。score设置为0.05来过滤掉大量背景bounding box。非最大抑制（NMS）设置为0.6。

## 2.2. 消除不一致

**RetinaNet (#A=1)**：只有一个anchor box的anchor-based RetinaNet。基本与FCOS相同，但是FCOS在$AP$性能上大大优于RetinaNet (#A=1)。FCOSv1 : RetinaNet = 37.1% : 32.5%。

**FCOS**：在v2版本上有一些对于v1的优化trick：将centerness移至Regression分支预测；使用GIoU损失函数；通过相应步长stride将Regression目标归一化。FCOSv2达到$AP = 37.8\%$。

**分析**

FCOSv1对比RetinaNet来说，有一些通用性的改进trick：

- head添加GroupNormalization (GN)；
- 使用GIoU Regression损失函数；
- 限制ground-truth box的正样本；
- 引入centerness；
- 添加一个可训练的标量。

而这些改进也可以应用于anchor-based的RetinaNet (#A=1)。因此这不是anchor-based和anchor-free的本质区别。

于是将这些FCOSv1的trick应用到ReinaNet中，以排除这些不一致。结果如Table 1所示。

Table 1: Analysis of implementation inconsistencies between RetinaNet and FCOS on MS COCO `minival` set. "#A=1" means there is one square anchor box per location.

| Inconsistency | FCOS | RetinaNet (#A=1) | | | | | |
|---|---|---|---|---|---|---|---|
| GroupNorm | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |
| GIoU Loss | ✓ | | | ✓ | ✓ | ✓ | ✓ |
| In GT Box | ✓ | | | | ✓ | ✓ | ✓ |
| Centerness | ✓ | | | | | ✓ | ✓ |
| Scalar | ✓ | | | | | | ✓ |
| AP (%) | 37.8 | 32.5 | 33.4 | 34.9 | 35.3 | 36.8 | 37.0 |

可见，全部应用FCOSv1 trick的RetinaNet (#A=1)依旧有$AP = 0.8\%$的差距。由此可以公平地探索他们的本质差异。

## 2.3. 本质差异

排除了通用性trick，现在只剩两个差异：

- 定义正负样本的划分方式；
- 从anchor box还是从anchor point开始Regression。

**定义正负样本的划分方式**

Table 2: Analysis of differences (%) between RetinaNet and FCOS on the MS COCO `minival` set.

| Classification \ Regression | Box | Point |
|---|---|---|
| Intersection over Union | 37.0 | 36.9 |
| Spatial and Scale Constraint | 37.8 | 37.8 |

RetinaNet从每层特征金字塔层级选择出anchor box与ground-truth $IoU > \theta_p$作为正样本，$IoU < \theta_n$作为负样本，然后忽略别的anchor box，对正负样本进行训练。即，RetinaNet利用IoU同时从spatial与scale上直接选择最终阳性。

而FCOS利用spatial与scale的限制 *来划分不同层的特征金字塔的anchor point。

如果将RetinaNet和FCOS选择的正负样本划定做一次交叉实验，实验结果如上图Table 2。即可发现，定义正负样本的划分方式对提高性能是本质差异之一。

**从anchor box还是从anchor point开始Regression**

依旧是从上图实验结果所示，到底是从anchor box还是anchor point开始Regression是一个无关紧要的点。

**结论**

定义正负样本的划分方式是本质区别。

# 3. 自适应训练样本选择ATSS

## 3.1. 介绍

关于超参数：anchor-based的IoU与之和anchor-free的scale范围是敏感的超参数，不同超参数设置将产生非常不同的结果。

ATSS可以不需要任何超参数就能根据物体的统计特征自动化分出正负样本。

ATSS具体算法流程图如下：

**Algorithm 1** Adaptive Training Sample Selection (ATSS)

**Input:**

$\mathcal{G}$ is a set of ground-truth boxes on the image
$\mathcal{L}$ is the number of feature pyramid levels
$\mathcal{A}_i$ is a set of anchor boxes from the $i_{th}$ pyramid levels
$\mathcal{A}$ is a set of all anchor boxes
$k$ is a quite robust hyperparameter with a default value of 9

**Output:**

$\mathcal{P}$ is a set of positive samples
$\mathcal{N}$ is a set of negative samples

1: **for** each ground-truth $g \in \mathcal{G}$ **do**
2: build an empty set for candidate positive samples of the ground-truth $g$: $\mathcal{C}_g \leftarrow \varnothing$;
3: **for** each level $i \in [1, \mathcal{L}]$ **do**
4: $\mathcal{S}_i \leftarrow$ select $k$ anchors from $A_i$ whose center are closest to the center of ground-truth $g$ based on L2 distance;
5: $\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;
6: **end for**
7: compute IoU between $\mathcal{C}_g$ and $g$: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;
8: compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;
9: compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;
10: compute IoU threshold for ground-truth $g$: $t_g = m_g + v_g$;
11: **for** each candidate $c \in \mathcal{C}_g$ **do**
12: **if** $IoU(c, g) \geq t_g$ and center of $c$ in $g$ **then**
13: $\mathcal{P} = \mathcal{P} \cup c$;
14: **end if**
15: **end for**
16: **end for**
17: $\mathcal{N} = \mathcal{A} - \mathcal{P}$;
18: **return** $\mathcal{P}, \mathcal{N}$;

**解读:**

- **(Algorithm1：Line 3-6)**

  对于图像上的每个ground-truth box $\mathcal{G}$，首先找到它在特征金字塔上**所有特征层**的候选positive样本 $\mathcal{C}_g$：于第 $i$ 层特征图上，从此层所有anchor box $\mathcal{A}_i$ 中选择 $k$ 个与 $\mathcal{G}$ 最接近的anchor box。$k$ 为超参数（几乎不影响结果）。

  假设有 $\mathcal{L}$ 层金字塔特征图，则一个ground-truth box $\mathcal{G}$ 有 $k \times \mathcal{L}$ 个候选positive样本 $\mathcal{C}_g$。

  > Question：
  >
  > L2 distance是什么意思?

- **(Algorithm1：Line 7-10)**

  Line 7：计算这些候选positive样本 $\mathcal{C}_g$ 与ground-truth box $\mathcal{G}$ 之间的IoU值，即 $\mathcal{D}_g$.

  Line 8-9：计算 $\mathcal{D}_g$ 的平均值 $m_g$ 和标准差 $v_g$.

  Line 10：可以得到 $\mathcal{G}$ 一个IoU阈值 $t_g = m_g + v_g$.

- **(Algorithm1：Line 11-15)**

  选择这些IoU大于等于阈值 $t_g$ 作为最终的正样本。

Line 12表明将正样本的选择限制在ground-truth box的中心区域。

此外，如果一个anchor box分配给多个ground-truth box，那么选择具有最高IoU的那个。

**Motivation：**

- **根据anchor box跟物体间中心距离选择候选anchor box。**距离越近，IoU越大。
- **使用 $m_g$ 和 $v_g$ 之和作为IoU的阈值 $t_g$ 。**一个对象的IoU平均 $m_g$ 是衡量预设anchor box对这个对象的合适程度。一个对象的IoU标准偏差 $v_g$ 是衡量哪些层适合检测这个物体的标准。使用 $m_g$ 和 $v_g$ 之和作为IoU的阈值 $t_g$ 可以根据对象的统计特征从适当的金字塔层中为每个对象自适应地选择足够的正样本。

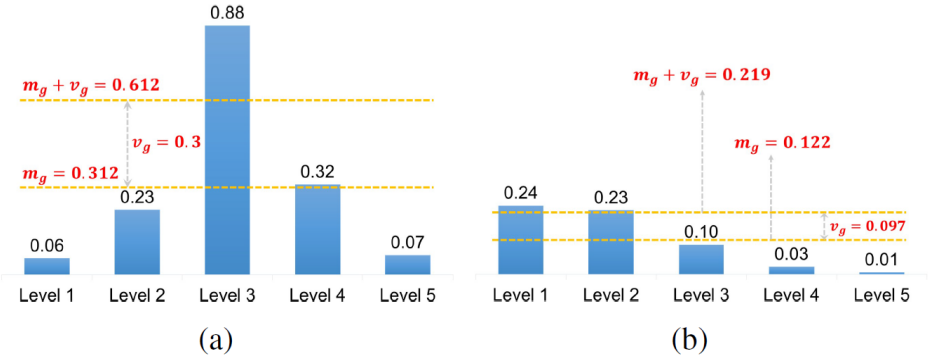


Figure 3: Illustration of ATSS. Each level has one candidate with its IoU. (a) A ground-truth with a high $m_g$ and a high $v_g$. (b) A ground-truth with a low $m_g$ and a low $v_g$.

**解释：**

如果每个level的IoU都高，那么平均 $m_g$就会高，这时候需要提高IoU阈值 $t_g$ 来适应均值都高的IoU们。描述这种情况的量便是IoU平均 $m_g$。

如果其中一个level的IoU比较高，那么说明这个level里面的基本都是质量高的anchor，所以优先选择这个level的anchor。相应的，IoU阈值 $t_g$ 也需要提高，以此来筛掉其他level的anchor。描述这种情况的量便是IoU标准偏差 $v_g$。

> Ref:
>
> [1] ATSS：论文与源码解读

- **保持不同对象之间的公平性。**RetinaNet和FCOS的策略往往对较大的物体有更多的正样本，导致不同物体之间的不公平。而ATSS保证了每个对象都有大约 $0.2 * k\mathcal{L}$ 的正样本。

## 3.2. 验证

Table 3: Verification of the proposed method (%) on the MS COCO `minival` set. ATSS and center sampling are the full version and the lite version of our proposed method.

| Method | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| RetinaNet (#A=1) | 37.0 | 55.1 | 39.9 | 21.4 | 41.2 | 48.6 |
| RetinaNet (#A=1) + ATSS | 39.3 | 57.5 | 42.8 | 24.3 | 43.3 | 51.3 |
| FCOS | 37.8 | 55.6 | 40.7 | 22.1 | 41.8 | 48.8 |
| FCOS + Center sampling | 38.6 | 57.4 | 41.4 | 22.3 | 42.5 | 49.8 |
| FCOS + ATSS | 39.2 | 57.3 | 42.4 | 22.7 | 43.1 | 51.5 |

## 3.3. 分析

**超参数 $k$**

进行了几个实验来研究超参数 $k$ 的robustness。实验结果如下所示：

Table 4: Analysis of different values of hyperparameter $k$ on the MS COCO `minival` set.

| $k$ | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 | 19 |
|---|---|---|---|---|---|---|---|---|---|
| AP (%) | 38.0 | 38.8 | 39.1 | 39.3 | 39.1 | 39.0 | 39.1 | 39.2 | 38.9 |

可见， $k$ 在7到17时基本没有变化。而太小的 $k$ 导致采样不完整导致统计不稳定，太大的 $k$ 导致夹杂较多的低质量anchor box致使性能稍微下降。但是整体的robustness是较好的， $k$ 作为超参数可以视为没有影响。

**anchor的大小**

Table 5: Analysis (%) of different anchor scales with fixed aspect ratio $1:1$ on the MS COCO `minival` set.

| Scale | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| 5 | 39.0 | 57.9 | 41.9 | 23.2 | 42.8 | 50.5 |
| 6 | 39.2 | 57.6 | 42.5 | 23.5 | 42.8 | 51.1 |
| 7 | 39.3 | 57.6 | 42.4 | 22.9 | 43.2 | 51.3 |
| 8 | 39.3 | 57.5 | 42.8 | 24.3 | 43.3 | 51.3 |
| 9 | 38.9 | 56.5 | 42.0 | 22.9 | 42.4 | 50.3 |

可见anchor大小变化具有robustness。

## 3.4. 对比

| Method | Data | Backbone | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|
| *anchor-based two-stage:* | | | | | | | | |
| MLKP [58] | trainval35 | ResNet-101 | 28.6 | 52.4 | 31.6 | 10.8 | 33.4 | 45.1 |
| R-FCN [9] | trainval | ResNet-101 | 29.9 | 51.9 | - | 10.8 | 32.8 | 45.0 |
| CoupleNet [74] | trainval | ResNet-101 | 34.4 | 54.8 | 37.2 | 13.4 | 38.1 | 50.8 |
| TDM [53] | trainval | Inception-ResNet-v2-TDM | 36.8 | 57.7 | 39.2 | 16.2 | 39.8 | 52.1 |
| Hu et al. [18] | trainval35k | ResNet-101 | 39.0 | 58.6 | 42.9 | - | - | - |
| DeepRegionlets [64] | trainval35k | ResNet-101 | 39.3 | 59.8 | - | 21.7 | 43.7 | 50.9 |
| FitnessNMS [57] | trainval | DeNet-101 | 39.5 | 58.0 | 42.6 | 18.9 | 43.5 | 54.1 |
| Gu et al. [15] | trainval35k | ResNet-101 | 39.9 | 63.1 | 43.1 | 22.2 | 43.4 | 51.6 |
| DetNet [31] | trainval35k | DetNet-59 | 40.3 | 62.1 | 43.8 | 23.6 | 42.6 | 50.0 |
| Soft-NMS [3] | trainval | ResNet-101 | 40.8 | 62.4 | 44.9 | 23.0 | 43.4 | 53.2 |
| SOD-MTGAN [1] | trainval35k | ResNet-101 | 41.4 | 63.2 | 45.4 | 24.7 | 44.2 | 52.6 |
| G-RMI [19] | trainval35k | Ensemble of Five Models | 41.6 | 61.9 | 45.4 | 23.9 | 43.5 | 54.9 |
| C-Mask RCNN [7] | trainval35k | ResNet-101 | 42.0 | 62.9 | 46.4 | 23.4 | 44.7 | 53.8 |
| Cascade R-CNN [5] | trainval35k | ResNet-101 | 42.8 | 62.1 | 46.3 | 23.7 | 45.5 | 55.2 |
| Revisiting RCNN [8] | trainval35k | ResNet-101+ResNet-152 | 43.1 | 66.1 | 47.3 | 25.8 | 45.9 | 55.3 |
| SNIP [54] | trainval35k | DPN-98 | 45.7 | 67.3 | 51.1 | 29.3 | 48.8 | 57.1 |
| *anchor-based one-stage:* | | | | | | | | |
| YOLOv2 [46] | trainval35k | DarkNet-19 | 21.6 | 44.0 | 19.2 | 5.0 | 22.4 | 35.5 |
| SSD512* [36] | trainval35k | VGG-16 | 28.8 | 48.5 | 30.3 | 10.9 | 31.8 | 43.5 |
| STDN513 [69] | trainval | DenseNet-169 | 31.8 | 51.0 | 33.6 | 14.4 | 36.1 | 43.4 |
| DES512 [68] | trainval35k | VGG-16 | 32.8 | 53.2 | 34.5 | 13.9 | 36.2 | 47.5 |
| DSSD513 [12] | trainval35k | ResNet-101 | 33.2 | 53.3 | 35.2 | 13.0 | 35.4 | 51.1 |
| RFB512-E [35] | trainval35k | VGG-16 | 34.4 | 55.7 | 36.4 | 17.6 | 37.0 | 47.6 |
| PFPNet-R512 [21] | trainval35k | VGG-16 | 35.2 | 57.6 | 37.9 | 18.7 | 38.6 | 45.9 |
| RefineDet512 [66] | trainval35k | ResNet-101 | 36.4 | 57.5 | 39.5 | 16.6 | 39.9 | 51.4 |
| RetinaNet [33] | trainval35k | ResNet-101 | 39.1 | 59.1 | 42.3 | 21.8 | 42.7 | 50.2 |
| *anchor-free keypoint-based:* | | | | | | | | |
| ExtremeNet [71] | trainval35k | Hourglass-104 | 40.2 | 55.5 | 43.2 | 20.4 | 43.2 | 53.1 |
| CornerNet [26] | trainval35k | Hourglass-104 | 40.5 | 56.5 | 43.1 | 19.4 | 42.7 | 53.9 |
| CenterNet-HG [70] | trainval35k | Hourglass-104 | 42.1 | 61.1 | 45.9 | 24.1 | 45.5 | 52.8 |
| Grid R-CNN [39] | trainval35k | ResNeXt-101 | 43.2 | 63.0 | 46.6 | 25.1 | 46.5 | 55.2 |
| CornerNet-Lite [27] | trainval35k | Hourglass-54 | 43.2 | - | - | 24.4 | 44.6 | 57.3 |
| CenterNet [11] | trainval35k | Hourglass-104 | 44.9 | 62.4 | 48.1 | 25.6 | 47.4 | 57.4 |
| RepPoints [65] | trainval35k | ResNet-101-DCN | 45.0 | 66.1 | 49.0 | 26.6 | 48.6 | 57.5 |
| *anchor-free center-based:* | | | | | | | | |
| GA-RPN [59] | trainval35k | ResNet-50 | 39.8 | 59.2 | 43.5 | 21.8 | 42.6 | 50.7 |
| FoveaBox [23] | trainval35k | ResNeXt-101 | 42.1 | 61.9 | 45.2 | 24.9 | 46.8 | 55.6 |
| FSAF [72] | trainval35k | ResNeXt-64x4d-101 | 42.9 | 63.8 | 46.3 | 26.6 | 46.2 | 52.7 |
| FCOS [56] | trainval35k | ResNeXt-64x4d-101 | 43.2 | 62.8 | 46.6 | 26.5 | 46.2 | 53.3 |
| *Ours:* | | | | | | | | |
| ATSS | trainval35k | ResNet-101 | 43.6 | 62.1 | 47.4 | 26.1 | 47.0 | 53.6 |
| ATSS | trainval35k | ResNeXt-32x8d-101 | 45.1 | 63.9 | 49.1 | 27.9 | 48.2 | 54.6 |
| ATSS | trainval35k | ResNeXt-64x4d-101 | 45.6 | 64.6 | 49.7 | 28.5 | 48.9 | 55.6 |
| ATSS | trainval35k | ResNet-101-DCN | 46.3 | 64.7 | 50.4 | 27.7 | 49.8 | 58.4 |
| ATSS | trainval35k | ResNeXt-32x8d-101-DCN | 47.7 | 66.6 | 52.1 | 29.3 | 50.8 | 59.7 |
| ATSS | trainval35k | ResNeXt-64x4d-101-DCN | 47.7 | 66.5 | 51.9 | 29.7 | 50.8 | 59.4 |
| ATSS (Multi-scale testing) | trainval35k | ResNeXt-32x8d-101-DCN | 50.6 | 68.6 | 56.1 | **33.6** | **52.9** | 62.2 |
| ATSS (Multi-scale testing) | trainval35k | ResNeXt-64x4d-101-DCN | **50.7** | **68.9** | **56.3** | 33.2 | **52.9** | **62.4** |

## 3.5. 讨论

还有一个点没有讨论：**每个位置平铺anchor的数量。**

原始的RetinaNet为每个位置平铺了9个anchor（3种比例的长宽比），标记为RetinaNet (#A=9)。在没有使用ATSS的情况下，RetinaNet (#A=9)要比RetinaNet (#A=1)有更好的性能。结果表明，在传统基于IoU的样本选择策略下，每个位置铺设更多的anchor是有效的。

但是使用ATSS后，铺设更多的anchor并没有多大用处，**需要进一步研究以发现期正确的作用。**