FCOS: Fully Convolutional One-Stage Object Detection

Zhi Tian, Chunhua Shen, Hao Chen, Tong He

摘要

我们提出了一种全卷积的单阶段目标检测器 (FCOS)，以类似于语义分割的每像素预测方式解决目标检测问题。几乎所有最先进的物体检测器，如 RetinaNet、SSD、YOLOv3 和 Faster R-CNN 都依赖于预定义的锚框。相比之下，我们提出的检测器 FCOS 没有锚框，也没有候选。通过消除预定义的锚框集合，FCOS 完全避免了与锚框相关的复杂计算，例如在训练过程中计算重叠。更重要的是，我们还避免了与锚框相关的所有超参数，最终检测性能通常对这些超参数非常敏感。凭借唯一的后处理——非极大值抑制（NMS），FCOS with ResNeXt-64x4d-101 在单模型单尺度测试的 AP 中达到了 44.7%，超越了之前的单阶段检测器，同时有着更简单的优势。我们第一次展示了一个更简单和灵活的检测框架，可以提高检测精度。我们希望提出的 FCOS 框架可以作为许多其他实例级任务的简单而强大的替代方案。代码位于：tinyurl.com/FCOSv1。

1 Introduction

目标检测是计算机视觉中一项基本但具有挑战性的任务，它需要算法为图像中的每个感兴趣的实例预测带有类别标签的边界框。目前所有主流检测器如 Faster R-CNN[3]、SSD[4] 和 YOLOv2、v3[8] 都依赖于一组预定义的锚框，长期以来一直认为使用锚框是检测器成功的关键。尽管取得了巨大的成功，但重要的是要注意基于锚的检测器有一些缺点：1) 如 [15, 24] 所示，检测性能对锚框的大小、长宽比和数量很敏感。例如，在 RetinaNet[7] 中，改变这些超参数最多可以在 COCO 基准测试上影响 4% 的 AP [16]。因此，需要在基于锚的检测器中仔细调整这些超参数。2）即使经过精心设计，由于锚框的尺度和纵横比保持固定，检测器在处理具有较大形状变化的候选对象时遇到困难，尤其是对于小对象。预定义的锚框也阻碍了检测器的泛

化能力，因为它们需要在具有不同对象大小或长宽比的新检测任务上重新设计。3）为了实现高召回率，一个基于锚的检测器需要在输入图像上密集放置锚框（例如，对于短边为 800 的图像，特征金字塔网络（FPN）[6] 中存在超过 180K 个锚框）。大多数这些锚框在训练期间被标记为负样本。过多的负样本加剧了训练中正负样本的不平衡。4) 锚框还涉及复杂的计算，例如使用 ground-truth 边界框计算 IoU 分数。

最近，全卷积网络 (FCN) [2] 在语义分割 [20、28、9、19]、深度估计 [17、31]、关键点检测 [3] 和计数 [2] 等密集预测任务中取得了巨大成功。作为高级视觉任务之一，目标检测可能是唯一一个偏离整洁的全卷积每像素预测框架的任务，主要是由于使用了锚框。提出一个问题是很自然的：*我们能否以简洁的每像素预测方式解决对象检测，例如，类似于语义分割的 FCN？* 因此，这些基本的视觉任务可以（几乎）统一在一个框架中。我们证明答案是肯定的。此外，我们首次证明，更简单的基于 FCN 的检测器比基于锚的检测器实现了更好的性能。

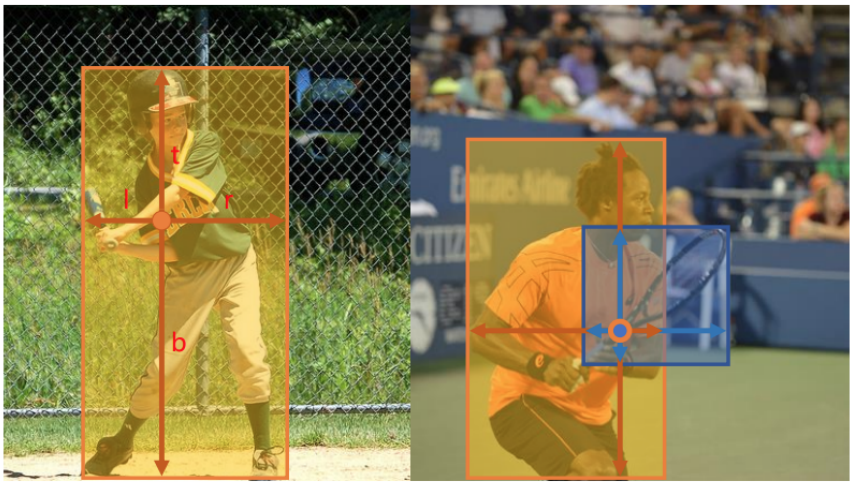

图 1: 如左图所示，FCOS 通过预测 4D 向量 (l, t, r, b) 对每个前景像素的边界框位置进行编码（在训练期间由 ground-truth 边界框信息监督）。右图显示，当一个位置在多个边界框中时，该位置应该回归哪个边界框可能不明确。

在文献中，一些工作尝试利用基于 FCN 的框架进行对象检测，例如 DenseBox[1]。具体来说，这些基于 FCN 的框架直接在每个特征图层级上

的每个空间位置预测 4D 向量和类别。如图1（左）所示，4D 向量描绘了从边界框的四个边到该位置的相对偏移量。这些框架类似于用于语义分割的 FCN，不同之处在于每个位置都需要回归 4D 连续向量。然而，为了处理不同大小的边界框，DenseBox[1] 将训练图像裁剪和调整到固定尺寸。因此 DenseBox 必须对图像金字塔进行检测，这违背了 FCN 一次计算所有卷积的理念。此外，更重要的是，这些方法主要用于特殊领域的物体检测，例如场景文本检测 [33, 10] 或人脸检测 [32, 12]，因为人们认为这些方法在应用于可能高度重叠的通用物体检测时效果不佳。如图 1（右）所示，高度重叠的边界框导致难以处理的歧义：不清楚 w.r.t. 重叠区域中的像素应该回归哪个边界框。

在续集中，我们仔细研究了这个问题，并表明使用 FPN 可以在很大程度上消除这种歧义。因此，我们的方法已经可以获得与那些传统的基于锚的检测器相当的检测精度。此外，我们观察到我们的方法可能会在远离目标对象中心的位置产生许多低质量的预测边界框。为了抑制这些低质量的检测，我们引入了一个新的“中心度”分支（只有一层）来预测像素与其相应边界框中心的偏差，如方程**??**中所定义。然后使用该分数降低检测到的低质量边界框的权重，并在 NMS 中合并检测结果。简单而有效的中心分支使得基于 FCN 的检测器在完全相同的训练和测试设置下优于基于锚的检测器。

这个新检测框架有如下优势。

- 检测现在与许多其他可以使用 FCN 解决的任务（例如语义分割）统一起来，从而更容易重用这些任务中的想法。
- 检测不再需要候选和锚框，这显着减少了设计参数的数量。设计参数通常需要启发式调整，并涉及许多技巧以实现良好的性能。因此，我们的新检测框架使检测器，尤其是其训练变得更加简单。
- 通过消除锚框，我们的新检测器完全避免了与锚框相关的复杂计算，例如训练期间锚框与真实框之间的 IOU 计算和匹配。从而相对于基于锚的对应检测器，实现更快的训练和测试以及更少的训练内存占用。
- 没有花里胡哨，我们在单阶段检测器中实现了最好的结果。我们还表明，所提出的 FCOS 可以用作两阶段检测器中的区域候选网络 (RPN)，并且可以实现比基于锚的 RPN 对应物明显更好的性能。鉴于更简单的无锚检测器的性能甚至更好，我们鼓励社区重新考虑锚框在目标检测中的必要性，它目前被认为是事实上的检测标准。

- 所提出的检测器可以立即以最少的修改扩展到其他视觉任务，包括实例分割和关键点检测。我们相信这种新方法可以成为许多实例预测问题的新基线。

2 Our Approach

在本节中，我们首先以每像素预测的方式重新制定目标检测。接下来，我们将展示我们如何利用多级预测来提高召回率并解决重叠边界框导致的歧义。最后，我们展示了我们提出的“centerness”分支，它有助于抑制检测到的低质量边界框并大幅提高整体性能。

2.1 全卷积单阶段物体检测器

令 $F_i \in \mathbb{R}^{H \times W \times C}$ 是主干 CNN 第 i 层的特征图，s 是到该层的总步长。输入图像的 ground-truth 边界框定义为 $\{B_i\}$，其中 $B_i = \left(x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)}, c^{(i)}\right) \in \mathbb{R}^4 \times \{1, 2, \ldots, C\}$。这里 $\left(x_0^{(i)}, y_0^{(i)}\right)$ 和 $\left(x_1^{(i)}, y_1^{(i)}\right)$ 表示边界框的左上角和右下角的坐标。$c^{(i)}$ 是边界框中的对象所属的类。C 是类的数量，对于 MS-COCO 数据集是 80。

对于特征图 F_i 上的每个位置 (x, y)，我们可以将其映射回输入图像的 $\left(\lfloor \frac{s}{2} \rfloor + xs, \lfloor \frac{s}{2} \rfloor + ys\right)$，它靠近位置的感受野中心 (x, y)。与基于锚点的检测器不同，它们将输入图像上的位置视为（多个）锚框的中心，并以这些锚框为参考回归目标边界框，我们直接在该位置回归目标边界框。换句话说，我们的检测器直接将位置视为训练样本，而不是基于锚的检测器中的锚框，这与用于语义分割的 FCN 相同 [2]。

具体来说，如果位置 (x, y) 落入任何 ground-truth 框，则将其视为正样本，并且该位置的类标签 c^* 是 ground-truth 框的类标签。否则它是一个负样本并且 c^*（背景类）。除了分类标签之外，我们还有一个 4D 实向量 $\mathbf{t}^* = (l^*, t^*, r^*, b^*)$ 作为该位置的回归目标。这里 l^*, t^*, r^* 和 b^* 是该位置到边界框四个边的距离，如图 1（左）所示。如果一个位置落入多个边界框，它被认为是一个有歧义的样本。我们简单地选择面积最小的边界框作为其回归目标。在下一节中，我们将展示通过多级预测，可以显着减少模糊样本的数量，因此它们几乎不会影响检测性能。形式上，如果位置 (x, y) 与边界

框 B_i 相关联，则该位置的训练回归目标可以表示为，

$$\begin{aligned} l^* &= x - x_0^{(i)}, \\ t^* &= y - y_0^{(i)}, \\ r^* &= x_1^{(i)} - x, \\ b^* &= y_1^{(i)} - y. \end{aligned} \tag{1}$$

值得注意的是，*FCOS* 可以利用尽可能多的前景样本来训练回归器。它与基于锚的检测器不同，后者只将与 ground-truth 框具有足够高 IOU 的锚框作为正样本。我们认为这可能是 FCOS 优于其基于锚的同行的原因之一。

网络输出。 对应于训练目标，我们网络的最后一层预测分类标签的 80D 向量 $\mathbf{p}$ 和 4D 向量 $\mathbf{t} = (l, t, r, b)$ 边界框坐标。按照 [7]，我们训练 C 个二元分类器，而不是训练多类分类器。与 [7] 类似，我们在主干网络的特征图之后添加了四个卷积层，分别用于分类和回归分支。此外，由于回归目标始终为正，因此我们在回归分支顶部使用 $\exp(x)$ 来将任何实数映射到 $(0, \infty)$。值得注意的是，FCOS 的网络输出变量比流行的每个位置有 9 个锚框的检测器 [7, 3] 少 9 倍，。

损失函数。 我们如下定义我们的训练损失函数:

$$\begin{aligned} L\left(\{\mathbf{p}_{x,y}\}, \{\mathbf{t}_{x,y}\}\right) =& \frac{1}{N_{\text{pos}}} \sum_{x,y} L_{\text{cls}}\left(\mathbf{p}_{x,y}, c^*_{x,y}\right) \\ &+ \frac{\lambda}{N_{\text{pos}}} \sum_{x,y} \mathbb{1}_{\{c^*_{x,y}>0\}} L_{\text{reg}}\left(\mathbf{t}_{x,y}, \mathbf{t}^*_{x,y}\right) \end{aligned} \tag{2}$$

其中 L_{cls} 是 [7] 中的焦点损失，L_{reg} 是 UnitBox[5] 中的 IOU 损失。N_{pos} 表示正样本的数量，λ 是 L_{reg} 的平衡权重，在本文中设置为 1。求和是在特征图 F_i 上的所有位置上计算的。$\mathbb{1}_{\{c^*_{x,y}>0\}}$ 是指示函数，如果 $c^*_i > 0$ 则为 1，否则为 0。

推理。 FCOS 的推理很简单。给定输入图像，我们通过网络将其转发并获得特征图 F_i 上每个位置的分类分数 $\mathbf{p}_{x,y}$ 和回归预测 $\mathbf{t}_{x,y}$。按照 [7]，我们选择 $p_{x,y} > 0.05$ 的位置作为正样本并反转方程1获得预测的边界框。

2.2 为 FCOS 使用 FPN 进行多层级预测

在这里，我们展示了如何通过 FPN[6] 的多级预测来解决 FCOS 的两个可能问题。1) CNN 中最终特征图的大步长（例如，16 倍）会导致相对较低的最佳召回率 (BPR——一个检测器所能达到的召回率的上界)。对于基于锚的检测器，由于大步长导致的低召回率可以通过降低正锚框所需的 IOU 分数在一定程度上得到补偿。对于 FCOS，乍一看，人们可能会认为 BPR 可能比基于锚的检测器低得多，因为由于步长较大，无法召回最终特征图上没有位置编码的对象。在这里，我们凭经验证明，即使有很大的步长，基于 FCN 的 FCOS 仍然能够产生很好的 BPR，甚至可以比 Detectron [7] 中在官方实现的基于锚的检测器 RetinaNet[7] 的 BPR 更好。所以 BPR 其实不是 FCOS 的问题。此外，通过多级 FPN 预测 [6]，可以进一步改进 BPR，以匹配基于锚的 RetinaNet 可以实现的最佳 BPR。2) 真实值框的重叠会导致难以处理的歧义，即重叠中的某个位置应该回归哪个边界框？这种模糊性导致基于 FCN 的检测器性能下降。在此工作中，我们表明多级预测可以极大地解决歧义，并且与基于锚的检测器相比，基于 FCN 的检测器可以获得同等甚至更好的性能。

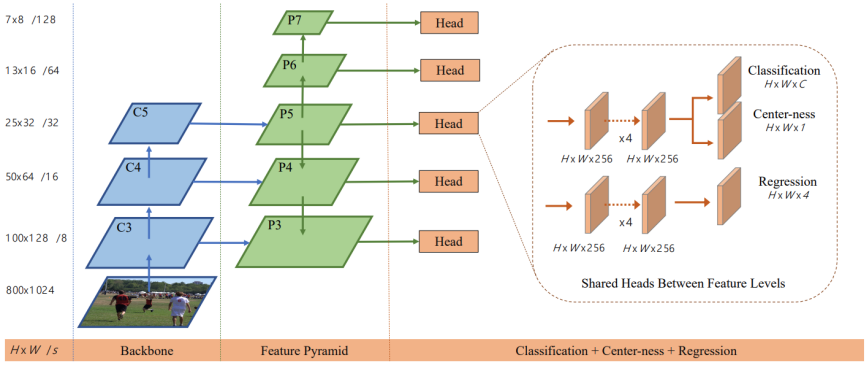

图 2: FCOS 的网络架构，其中 C3、C4、C5 表示骨干网络的特征图，P3 到 P7 是用于最终预测的特征层级。$H \times W$ 是特征图的高度和宽度。'/s' ($s = 8, 16, \ldots, 128$) 是特征图对输入图像层级的下采样率。例子中所有数字都是使用 800×1024 的输入计算的。

遵循 FPN[6]，我们在不同层级的特征图上检测不同大小的对象。具体来说，我们使用了五级特征图 P_3, P_4, P_5, P_6, P_7。P_3, P_4, P_5 由主干 CNN 的

特征图 C_3, C_4, C_5 通过一个 1×1 卷积层，以及在 [6] 中自顶向下的连接产生，如图2所示。P_6, P_7 分别通过在 P_5, P_6 上应用一个步幅为 2 的卷积层产生。因此，特征级别 P_3, P_4, P_5, P_6 和 P_7 的步长分别为 8、16、32、64 和 128。

Unlike anchor-based detectors, which assign anchor boxes with different sizes to different feature levels, we directly limit the range of bounding box regression for each level. More specifically, we firstly compute the regression targets l∗, t∗, r∗ and b∗ for each location on all feature levels. Next, if a location satisfies max(l∗; t∗; r∗; b∗) > mi or max(l∗, t∗; r∗; b∗) < mi−1, it is set as a negative sample and is thus not required to regress a bounding box anymore. Here mi is the maximum distance that feature level i needs to regress. In this work, m2, m3, m4, m5, m6 and m7 are set as 0, 64, 128, 256, 512 and 1, respectively. Since objects with different sizes are assigned to different feature levels and most overlapping happens between objects with considerably different sizes. If a location, even with multi-level prediction used, is still assigned to more than one ground-truth boxes, we simply choose the groundtruth box with minimal area as its target. As shown in our experiments, the multi-level prediction can largely alleviate the aforementioned ambiguity and improve the FCN-based detector to the same level of anchor-based ones.

References

[1] Lichao Huang et al. “Densebox: Unifying landmark localization with end to end object detection”. In: *arXiv preprint arXiv:1509.04874* (2015).

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[3] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

[4] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[5] Jiahui Yu et al. “Unitbox: An advanced object detection network”. In: *Proceedings of the 24th ACM international conference on Multimedia*. 2016, pp. 516–520.

[6] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[7] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[8] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).