Deformable DETR: Deformable Transformers for End-to-End Object Detection

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai
arXiv:2010.04159 [cs.CV] (or arXiv:2010.04159v4 [cs.CV])

智能网络与优化实验室

2023/5/3

1 Introduction

Motion

DETR 的问题

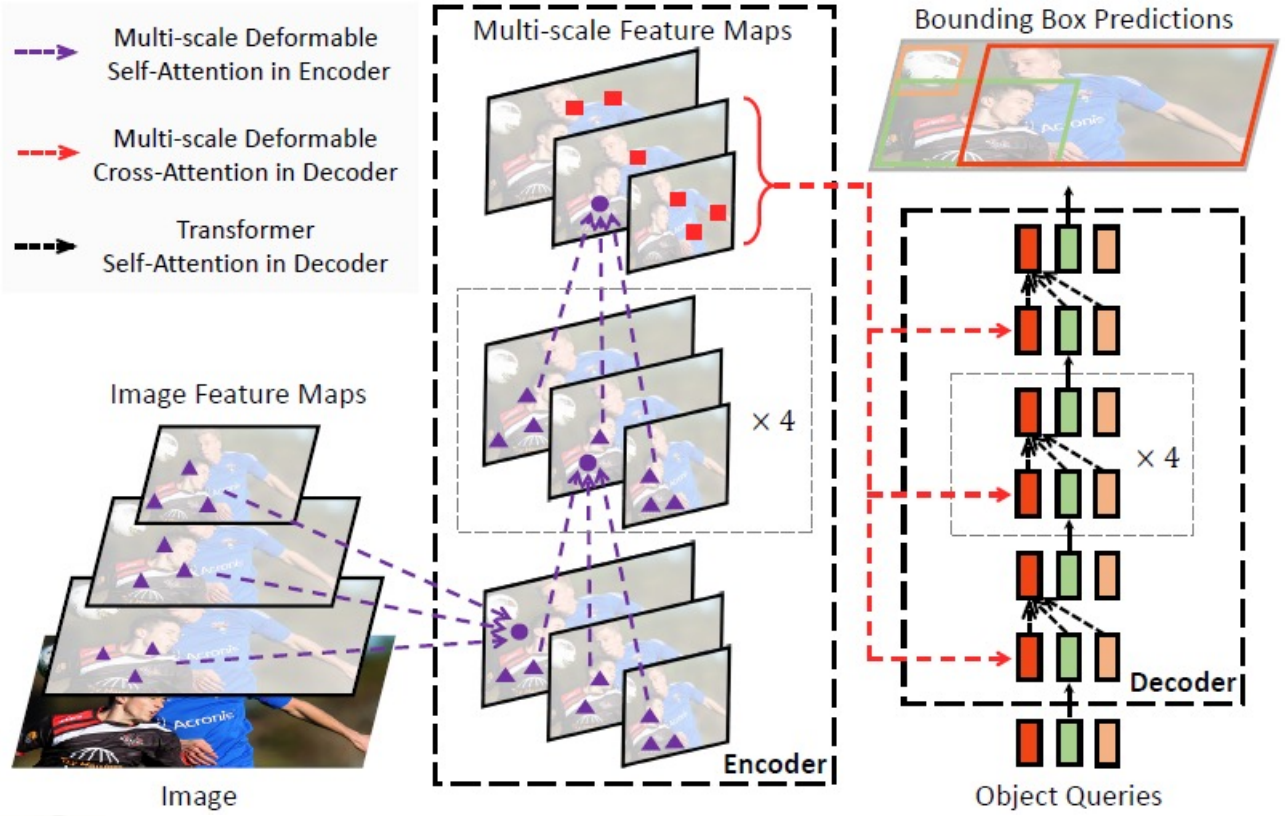
- 相比与现有的目标检测器，DETR 需要更长的训练时间来收敛。例如 COCO 数据集上的收敛速度较 Faster R-CNN 慢 10 到 20 倍。
- DETR 对小目标的检测性能相对较低。目标检测器通常利用多尺度特征，从高分辨率特征图中检测小物体。但是，高分辨率的 feature maps 将大幅提高 DETR 的复杂度。

问题的成因分析

- 主要原因：transformer 在处理图像特征图时存在缺陷。
- 初始化时，注意力模块对特征图中的所有像素施加基本一致的注意力权重。为了让注意力集中在稀疏而有意义的位置，需要长时间的训练。
- transformer 编码器的注意力权重计算是基于像素数的二次复杂度。因此，处理高分辨率特征图的计算和内存复杂度较高。

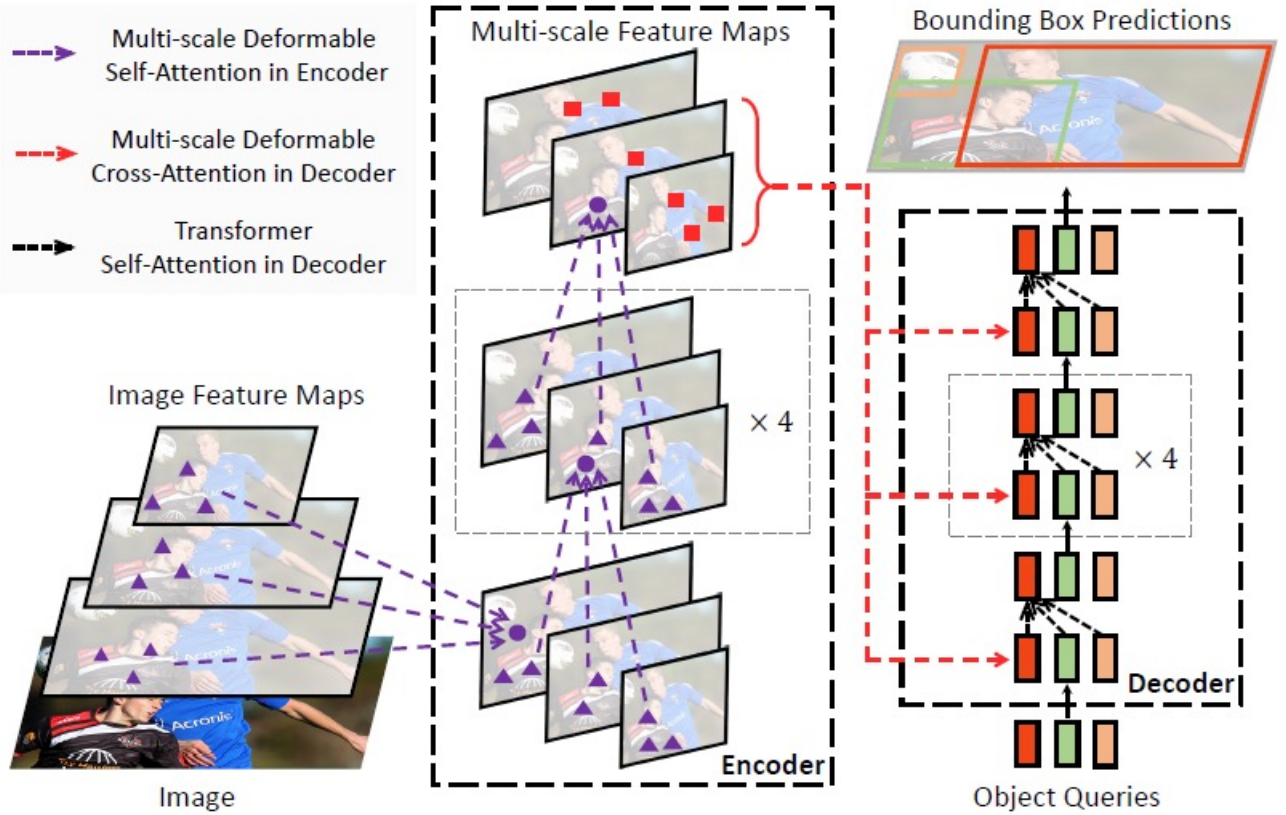
改进方案

- 提出 Deformable DETR，可以缓解 DETR 收敛速度慢和复杂度高的问题。它融合了可变形卷积良好的稀疏空间采样能力和transformer的强大关系建模能力。
- 提出可变形注意模块，仅关注特征图上一小部分关键的采样点。并且该模块可以自然地扩展到聚合多尺度特征，而无需 FPN 的帮助。在 Deformable DETR 中，使用（多尺度）可变形注意模块替换原有 transformer 的注意力模块。



改进方案

- 如图，由于 transformer 编码器中的 self-attention 和 encoder-decoder 连接的 cross attention 涉及到像素级的相关性计算，复杂度比较高，这两个 attention 相应的被替换为稀疏的可变形注意力模块，decoder 中的 query self-attention 复杂度不高，沿用原来的 transformer 的 attention 模块。



2 Methodology

Transformer 和 DETR

Transformers 中的多头注意力机制，每个 head 单独刻画 queries 与 keys 之间的注意力加权，然后通过权重线性加权获得最终的输出，计算如下：

$$\mathrm{MultiHeadAttn}(z_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k \in \Omega_k} A_{mqk} \cdot W'_m x_k \right]$$

- $q \in \Omega_q$ 表示某个 query 元素，特征表示为 $z_q \in \mathbb{R}^C$；$k \in \Omega_k$ 表示某个 key 元素，特征表示为 $x_k \in \mathbb{R}^C$，C 为特征维度
- Ω_q 和 Ω_k 为所有 key 构成的集合
- $W'_m \in \mathbb{R}^{C_v \times C}$ 和 $W_m \in \mathbb{R}^{C \times C_v}$ 分别是可学习的权重，$C_v = C/M$，M 表示 attention head 的数量
- $A_{mqk} \propto \exp \frac{z_q^T U_m^T V_m x_k}{\sqrt{C_v}}$ 是第 k 个元素的注意力权重，$\sum_{k \in \Omega_k} A_{mqk} = 1$，$U_m, V_m \in \mathbb{R}^{C_v \times C}$ 是可学习的权重
- $W'_m x_k$ 是对 key 元素的编码
- z_q, x_k 一般是内容特征和位置编码的聚合

Transformer 和 DETR

- **训练时间长**：以 N_q 和 N_k 表示 query 和 key 的数量，初始 attention 的权重归一化后约为 $1/N_k$，在图像中 N_k 一般很大，导致 N_k 对输入特征产生模糊的梯度，因此需要长时间的训练。
- **复杂度分析**：多头注意力机制的计算复杂度为 $O(N_qC^2 + N_kC^2 + N_qN_kC)$，在图像中 N_q 和 N_k 远大于 C，因此计算复杂度主要为第三项 $O(N_qN_kC)$。

transformer 中三个 attention 模块的复杂度如表：

模块	复杂度
encoder self-attention	$O(H^2W^2C)$
decoder cross-attention	$O(HWC^2) + NHWC$
decoder self-attention	$O(2NC^2 + N^2C)$

Deformable Attention Module

Deformable DETR Attention 的注意力机制为：

$$\mathrm{DeformAttn}(z_q, p_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} \cdot W'_m x(p_q + \Delta p_{mqk}) \right]$$

- 与 DETR 相比，将 key 的范围从 Ω_k 限定到了 K 个采样点， $\sum_{k=1}^{K} A_{mqk} = 1$，每个采样点都有一个预测的采样点 offset（ Δp_{mqk} ），以达到自适应位置的采样点。
- $p_q + \Delta p_{mqk}$ 通常是小数，从特征图上索引特征时采用双线性插值的方式。
- Δp_{mqk} 和 A_{mqk} 是通过对 query 特征 z_q 进行线性映射得到的。在实现中，query 特征 z_q 被通过一个 $3MK$ 通道的线性投影运算操作，其中前 $2MK$ 通道预测采样偏移量 Δp_{mqk}，其余 MK 通道被提供给 softmax 获得注意力权重 A_{mqk}

Multi-scale Deformable Attention Module

给定 $\{x_l\}_{l=1}^{L}$ 表示输入多尺度特征图（multi-scale feature map），则注意力计算为：

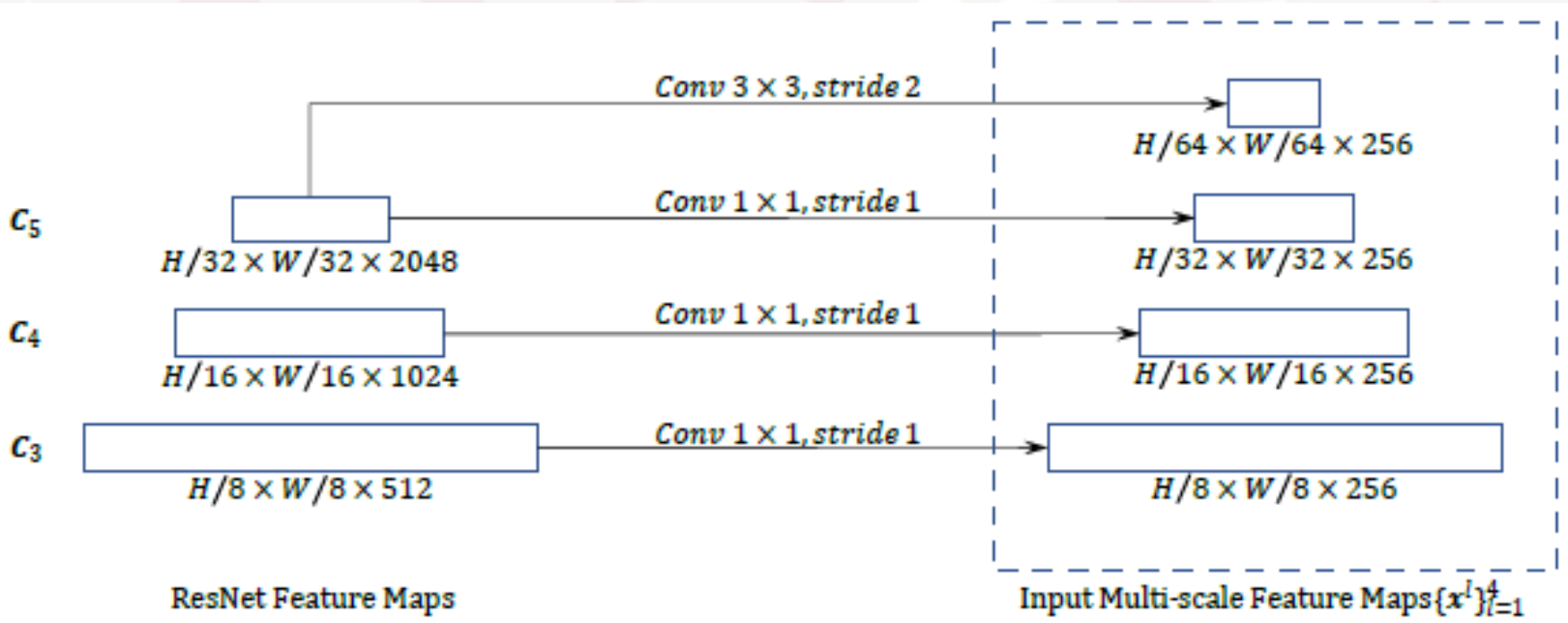
$$\mathrm{MSDeformAttn}\left(z_q, \hat{p}_q, \{x_l\}_{l=1}^{L}\right) = \sum_{m=1}^{M} W_m \left[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} \cdot W'_m x^l\left(\phi_l\left(\hat{p}_q\right) + \Delta p_{mlqk}\right)\right]$$

其中 l 表示不同的尺度层，$\phi_l(\cdot)$ 表示将对应位置映射到第 l 层，W_m 每一层共享， $\hat{p}_q$ 头上的尖角代表归一化到 $[0,1]$。

此时，采样点变成了从多尺度 feature map 中每层选取 K 个采样点，共计 LK 个点，并且有 $\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} = 1$

Deformable Transformer Encoder

- encoder 输入和输出均是多尺度特征图。
- encoder 输入的多尺度特征共有 4 层（下图），其中 C3~C5 来自 ResNet 的最后 3 个 stage，下采样率对应为 8，16，32，C6 由 C5 经过步长为 2 的 3×3 卷积得到。
- 为了引入位置信息，query 和 key 通常会加上固定的 position encoding。
- 为了引入每个像素的尺度（level）信息，给每个像素学习一个尺度 embedding $\{e_l\}_{l=1}^{L}$。



Deformable Transformer Decoder

- decoder 中包含 cross-attention 和 self-attention，两者的注意力 query 均是 object query，decoder cross-attention 的 key 是 encoder 输出的多尺度特征图，decoder self-attention 仍是 object query。

- 对于每个 object query，都会预测其对应点的 2-d 归一化坐标 $\hat{p}_q$，由线性映射层+sigmoid 激活函数学习得到。

- 由于多尺度可变形注意力模块提取的是参考点周围的图像特征，所以让检测头预测的 bbox 为相对参考点的相对偏移量，从而进一步降低优化难度。参考点作为初始的 bbox center 预测，bbox head 预测相对参考点的偏移量，加速训练的收敛。

Deformable DETR 的其他变种

1. Iterative Bounding Box Refinement

针对于多个 decoder，每个 decoder 的输入是前一层的 decoder 的输出。

2. Two-Stage Deformable DETR

DETR 中的 query 是随机初始化的，而两阶段方式则是由 Deformable DETR 的变种生成初始的候选 query。在第一阶段，移除 Deformable DETR 中的 decoder 模块，仅使用 encoder 模块，每个像素位置都作为 query，直接预测 box，然后选择 score 最大的 N 个 box 作为 region proposal 输入下一阶段。

3 Experiment

实验设置

- 多头注意力机制 $M = 8$
- 不同的 scale $K = 4$
- encoder 共享参数
- 其余超参数的设置参考 DETR，其中 Focal Loss 权重变为 2，query 个数从 100 增为 300
- 使用 Adam 优化器，取学习率为 2×10^{-4}，$\beta_1 = 0.9$，$\beta_2 = 0.999$，权重衰减为 10^{-4}

与 DETR 的对比

DC5 表示对 ResNet 网络最后一层 stride 以及空洞卷积保持分辨率的修改。DC5+ 表示与当前的训练设置相同设置 DETR 模型的训练结果。

可以发现 Deformable DETR 的最大贡献点在于训练周期的大幅缩短以及对小目标精度的提升。

Table 1: Comparision of Deformable DETR with DETR on COCO 2017 val set. DETR-DC5+ denotes DETR-DC5 with Focal Loss and 300 object queries.

Method	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	Training GPU hours	Inference FPS
Faster R-CNN + FPN	109	42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	380	26
DETR	500	42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	2000	28
DETR-DC5	500	43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	7000	12
DETR-DC5	50	35.3	55.7	36.8	15.2	37.5	53.6	41M	187G	700	12
DETR-DC5+	50	36.2	57.0	37.4	16.3	39.2	53.9	41M	187G	700	12
Deformable DETR	50	43.8	62.6	47.7	26.4	47.1	58.0	40M	173G	325	19
+ iterative bounding box refinement	50	45.4	64.7	49.0	26.8	48.3	61.7	40M	173G	325	19
++ two-stage Deformable DETR	50	46.2	65.2	50.0	28.8	49.2	61.7	40M	173G	340	19

与 DETR 的对比

DC5 表示对 ResNet 网络最后一层 stride 以及空洞卷积保持分辨率的修改。DC5+ 表示与当前的训练设置相同设置 DETR 模型的训练结果。

可以发现 Deformable DETR 的最大贡献点在于训练周期的大幅缩短以及对小目标精度的提升。

Table 1: Comparision of Deformable DETR with DETR on COCO 2017 val set. DETR-DC5+ denotes DETR-DC5 with Focal Loss and 300 object queries.

Method	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	Training GPU hours	Inference FPS
Faster R-CNN + FPN	109	42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	380	26
DETR	500	42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	2000	28
DETR-DC5	500	43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	7000	12
DETR-DC5	50	35.3	55.7	36.8	15.2	37.5	53.6	41M	187G	700	12
DETR-DC5+	50	36.2	57.0	37.4	16.3	39.2	53.9	41M	187G	700	12
Deformable DETR	50	43.8	62.6	47.7	26.4	47.1	58.0	40M	173G	325	19
+ iterative bounding box refinement	50	45.4	64.7	49.0	26.8	48.3	61.7	40M	173G	325	19
++ two-stage Deformable DETR	50	46.2	65.2	50.0	28.8	49.2	61.7	40M	173G	340	19

与 SOTA 的对比

Table 3: Comparison of Deformable DETR with state-of-the-art methods on COCO 2017 test-dev set. "TTA" indicates test-time augmentations including horizontal flip and multi-scale testing.

Method	Backbone	TTA	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS (Tian et al., 2019)	ResNeXt-101		44.7	64.1	48.4	27.6	47.5	55.6
ATSS (Zhang et al., 2020)	ResNeXt-101 + DCN	✓	50.7	68.9	56.3	33.2	52.9	62.4
TSD (Song et al., 2020)	SENet154 + DCN	✓	51.2	71.9	56.0	33.8	54.8	64.2
EfficientDet-D7 (Tan et al., 2020)	EfficientNet-B6		52.2	71.4	56.3	-	-	-
Deformable DETR	ResNet-50		46.9	66.4	50.8	27.7	49.7	59.9
Deformable DETR	ResNet-101		48.7	68.1	52.9	29.1	51.5	62.0
Deformable DETR	ResNeXt-101		49.0	68.5	53.2	29.7	51.7	62.8
Deformable DETR	ResNeXt-101 + DCN		50.1	69.7	54.6	30.6	52.8	64.7
Deformable DETR	ResNeXt-101 + DCN	✓	52.3	71.9	58.1	34.4	54.4	65.6

4 References

参考

- https://niecongchong.blog.csdn.net/article/details/118693939
- https://zhuanlan.zhihu.com/p/372116181
- https://www.jianshu.com/p/8524abf10018

源码

- https://github.com/fundamentalvision/Deformable-DETR

谢谢

Thank You

THANKS