DOI:10.19651/j.cnki.emt.2210400

基于预过滤注意力的 Transformer 物体检测

王 琪 赵文仓
(青岛科技大学自动化与电子工程学院 青岛 266061)

摘 要: 近几年提出的基于 Transformer 的目标检测器简化了模型结构,展现出具有竞争力的性能。然而,由于 Transformer 注意力模块处理特征图的方式,大部分模型存在收敛速度慢和小物体检测效果差的问题。为了解决这些问题,本研究提出了基于预过滤注意力模块的 Transformer 检测模型,该模块以目标点为参照,提取目标点附近部分特征点进行交互,节省训练时长并提高检测精度。同时在该模块中融入新提出的一种有向相对位置编码,弥补因模块权重计算导致的相对位置信息缺失,提供精确的位置信息,更有利于模型对小物体的检测。在 COCO 2017 数据集上的实验表明我们的模型可以将训练时长缩短近 10 倍,并获得更好的性能,特别是在小物体检测上精度达到了 26.8 APs。

关键词: 目标检测;Transformer;位置编码;注意力机制
中图分类号: TP391.9 **文献标识码:** A **国家标准学科分类代码:** 520.60

Object detection based on Transformer with prefiltered attention

Wang Qi Zhao Wencang
(College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Transformer-based target detectors proposed in recent years have simplified the model structure and demonstrated competitive performance. However, most of the models suffer from slow convergence and poor detection of small objects due to the way the Transformer attention module handles feature maps. To address these issues, this study proposes a Transformer detection model based on a pre-filtered attention module. Using the target point as reference, the module only samples a part of the feature points near the target point, which saves training time and improves detection accuracy. A newly defined directional relative position encoding is also integrated in the module. The encoding compensates for the lack of relative position information in the module due to the weight calculation that is more helpful for the detection of small objects. Experiments on the COCO 2017 dataset show that our model reduces the training time by a factor of 10 and improves the detection accuracy, especially on small object detection by 26.8 APs.

Keywords: object detection;Transformer;location coding;attention mechanism

0 引 言

现有的目标检测模型主要基于卷积神经网络(convolutional neural network,CNN)[1],大多使用 Region Proposal[2]、Anchor[3]、Window Center[4]、非极大值抑制(non-maximum suppression,NMS)[5]等间接技术,借助回归或分类等方法检测目标的位置和类别,但这些方法在一定程度上都借助了人工辅助,导致模型中间步骤冗余,影响模型的最终效果。

随着 Transformer[6]在计算机视觉领域的发展,研究者们开始将其引入到目标检测中,与 CNN 结合,把检测任务描述为集合预测问题,结合二分匹配进行检测。Transformer 的引入降低了检测模型对人工组件的依赖,简化模型架构实现端到端检测,这使得 Transformer 成为目标检测中一个新的研究方向。然而 Transformer 在处理图像时仍存在不足,一方面 Transformer 中注意力模块由全局过渡到局部的处理方式,导致模型收敛速度变慢。另一方面,Transformer 在处理高分辨率的特征时会伴随极高的计算量,这对检测小目标很不友好。此外,Transformer 需结合位置编码以学习序列顺序,现有的三角函数式位置编码在提供绝对位置的同时,还在内部计算过程中产生隐式相对位置信息,以此来更好的学习序列位

收稿日期:2022-06-21

置，但这种相对位置信息没有方向性，还会因后续计算或注意力模块改变而消失，同样影响模型检测小目标的精度。

为了解决 Transformer 的不足，我们提出了一种基于预过滤注意力模块的 Transformer 检测模型，模型通过改进 Transformer 注意力模块处理特征图的方式，实现了一种轻量级预过滤注意力机制，解决 Transformer 检测模型收敛速度慢以及小目标检测效果差的问题。模型注意力模块舍弃特征像素与全局特征交互的想法，对 key 进行稀疏化提取，可以看作是一种聚焦关键特征的过滤器，减少计算成本提高收敛速度和检测精度。同时在该注意力模块中融入一种新定义的二维有向相对位置编码以更好的学习位置信息，该有向相对位置编码可以看作是注意力机制的扩展，弥补注意力模块中因权重计算导致的隐式相对位置信息缺失，有利于提高检测小物体的准确度。此外，与目前大多数检测器不同，该模块不需借助特征金字塔网络（feature pyramid networks，FPN）[7] 就可以在多尺度层面上实现特征跨层融合。

在 COCO 2017[8] 数据集上进行大量实验来评估基于预过滤注意力模块的 Transformer 检测模型，实验表明我们的模块进一步提高了 Transformer 检测模型的性能，可以在大幅度缩减训练时长近 10 倍的情况下获得与基线相当甚至更好的性能，尤其在小物体检测上精确度最大提高了 6.3 APs。另外，我们采用的有向相对位置编码的方法也有效的提高模型的精测精度，证明了目标检测中相对位置编码的有效性。

1　相关工作

1.1　基于 Transformer 的目标检测

现有检测器大多基于 CNN 算法，可分为一阶段检测算法[9] 和两阶段检测算法[2]，尽管这些模型都有过人之处，但大部分依赖于 NMS 等手工组件，使模型性能大打折扣。近几年，Carion 等[10] 提出了基于 Transformer 的端到端目标检测器（detection transformer，DETR），模型结合 CNN 和 Transformer 将检测任务描述为集合预测问题，结合二分匹配直接进行检测。DETR 的成功使 Transformer 成为目标检测领域新兴的研究方向，针对 Transformer 的不同问题，各种检测器开始层出不穷。从数据的角度，UP-DETR[11] 提出了一个前置任务，无监督地预训练 DETR 的 Transformer，收敛速度得到提升但在小物体检测上仍不理想。另外，Dynamic DETR[12] 设计了一种动态解码器，能够以粗粒度到细粒度的方式关注兴趣区域，降低学习难度，但模型中的 Transformer 需要从头开始训练，这限制了检测模型的鲁棒性和泛化能力。而 DN-DETR[13] 从训练方法出发，提出利用去噪训练解决 DETR 二分匹配不稳定的问题。本研究从注意力模块计算方式出发，提出了预过滤注意力模块，借助可变形卷积稀疏空间位置的能力，只关注一部分采样点并进行交互，在一定程度上降低了注意力模块计算导致的各种消耗。

1.2　位置编码

Transformer 自被提出以来发展迅速，逐渐开始应用到视觉领域，不同于循环神经网络（recurrent neural networks，RNN）[14] 等模型，Transformer 自身无法学习序列的顺序，然而不论自然语言处理还是视觉任务中，位置信息都是必不可少的。针对这个问题，最初研究者们提出了绝对位置编码，将输入位置从 1 到最大序列长度依次编码，此时位置向量产生在自注意力模块外，与输入嵌入相加再送入 Transformer。与绝对位置编码不同，后来提出的相对位置编码[15] 是在注意力模块计算过程中考虑输入间的相对距离，为了充分利用位置信息，Huang 等[16] 又提出了一种同时考虑 query、key 和相对位置交互的方法。Dai 等[17] 从绝对编码的表达式出发，为 query 添加偏置项，并使用正弦公式对相对位置编码。此外，Ramachandran 等[18] 提出了一种专门针对二维图像的编码方法，将相对编码分为水平和垂直两个方向分别进行编码建模，最后再进行拼接。本研究在使用现有三角函数式位置编码的同时，又在注意力机制中定义了一种有向相对位置编码以弥补相对位置信息的缺失，采用类似偏差模式的方法将编码融入到注意力机制中，给模型提供更精确的输入位置信息。

2　方法与架构

目前基于 Transformer 的检测模型通过结合 CNN 和 Transformer 将检测任务转化为集合预测，利用二分匹配进行端到端检测，虽然提高了模型性能，但仍有收敛慢和小目标检测精度差的问题，我们分析这些问题主要与 Transformer 注意力模块有关。由于注意力模块遍历特征图全局的注意力，模型需要较长的训练周期学习关注有意义的位置。其次，现有检测器通常使用多尺度特征检测不同大小的物体，但 Transformer 处理高分辨率特征时会带来计算量的激增，不利于小目标的检测。此外，为了学习输入序列顺序，常用三角函数式位置编码在 Transformer 输入中添加绝对位置，同时借助注意力模块的点积计算在内部产生隐式相对位置信息，但这种相对位置信息没有方向性，还容易在注意力后续计算中消失，同样影响小目标检测精度。

为了解决这些问题，提出了基于预过滤注意力模块的 Transformer 检测模型，其注意力模块在提取 key 时进行稀疏化处理，只采样目标点附近部分像素作为高相关性 key 与 query 进行交互，减少模块全局建模的消耗，提高收敛速度与精度。此外，由于对注意力机制的改动，注意力权重的获取不再依赖 query 和 key 的点积操作，导致原本借助点积产生的隐式相对位置信息缺失，因此我们又新定义了一种有向相对位置编码，采用类似偏差模式的方法将该编码融入到注意力机制中，补充相对位置信息以更好的检测小物体。本研究中应用的是多尺度预过滤注意力模块，模型架构主要包括 CNN 特征提取网络、基于注意力的

Transformer 和用于最后检测的前馈网络三部分,具体架构如图 1 所示。

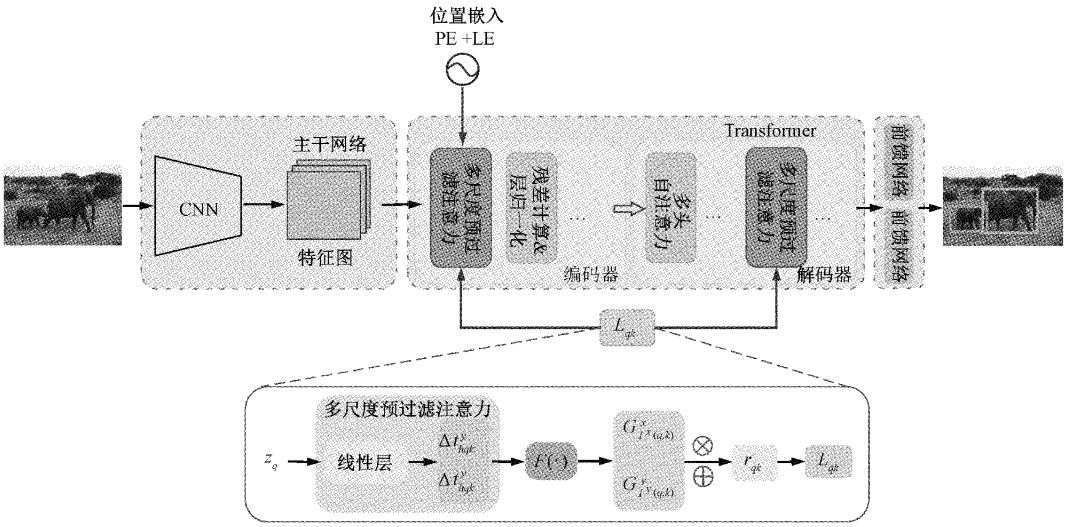

图 1　检测模型总体架构

2.1　CNN 特征提取网络

由于模型可以用于处理多尺度特征,因此我们以传统的 ResNet-50[19]网络作为模型的核心特征提取部分,从中提取不同层次的特征图层处理后作为模型的多尺度输入,并且不需要使用 FPN,研究提出的预过滤注意力模块就可以实现多尺度特征的跨层融合。为了增强特征提取的能力,首先我们将 ResNet 在阶段 3、4 中生成的特征图记作 C3、C4,然后替换 ResNet 最后一阶段的卷积为可变形卷积,得到可变形的特征图 C5,最后一层特征图由 C5 特征层经 3×3 的卷积获得,具体操作方法如图 2 所示。

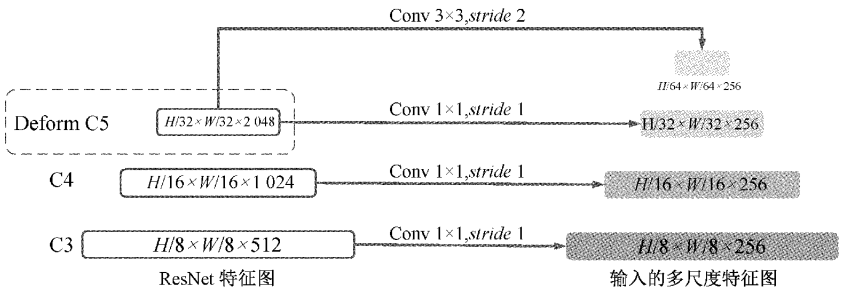

图 2　构建多尺度特征图

2.2　基于预过滤注意力的 Transformer

1)预过滤注意力模块

假设模块特征图输入是 $x \in R^{C\times H\times W}$,则预过滤注意力表示为:

$$Attn(z_q,t_q,x)=\sum_{h=1}^{H}W_h\left[\sum_{k=1}^{K}\alpha_{hqk}\cdot TW^V\right] \tag{1}$$

$$T=f(t_q+\Delta t_{hqk}) \tag{2}$$

其中,z_q 代表由输入线性变换得到的查询特征 query,t_q 为 z_q 对应的位置坐标,这里称作目标点。k 指采样的 key,h 代表注意力头数,每个 query 在 H 个注意力头中分别采样 K 个位置,且只和采样特征交互。$\boldsymbol{W}_h$ 与 $\boldsymbol{W}^V$ 均为可学习的权重矩阵。Δt_{hqk} 表示第 h 个注意力头中第 k 个采样点基于目标点的偏移量,由 z_q 通过线性层获取。采样点的 value 值 T 是由最终的采样点坐标(偏移量+目标点)应用双线性插值 $f(t_q+\Delta t_{hqk})$ 计算得到。α_{hqk} 表示第 h 个注意力头中第 k 个采样点的注意力权重,这里注意力权重的获取不再依赖 query 和 key 间的点积计算,而是同偏移量一样,由 query 经全连接层线性投影得到,由于这种方式破坏了原始 Transformer 中三角函数式位置编码通过点积操作得到的隐式相对位置信息,因此我们在计算权重过程中融入了一种新定义的二维有向相对位置编码,以补充缺失的相对位置信息提高小物体检测的精度,所以这里 α_{hqk} 包含两部分:在第 h 个注意力头中,首先是通过查询特征 z_q 线性投影得到的注意力分数 e'_{qk},其次是对应计算得到的相对位置编码 L_{qk},如式(3),最后两者做和输入到 Softmax 层得到注意力分布 α_{hqk}。预过滤注意力模块具体架构如图 3 所示。

2)相对位置编码的计算

(1)相对位置模式

根据编码是否独立于输入将相对位置分为偏差和上

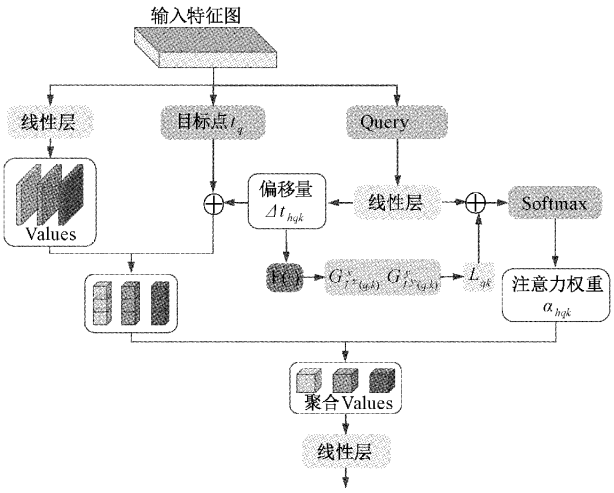

图 3　预过滤注意力模块图示

下文两种模式，本研究使用与偏差模式相类似的方法计算相对位置编码，编码独立于输入嵌入，如式(3)，这里只列举单个注意力头的计算，多头注意力的计算再进行集成即可。

$$e_{qk} = e'_{qk} + L_{qk} \tag{3}$$

$$\alpha_{qk} = soft\max(e_{qk}) = soft\max(e'_{qk} + L_{qk}) \tag{4}$$

在该偏差模式下有：

$$L_{qk} = r_{qk} \tag{5}$$

其中，e_{qk} 为模型中最终的注意力分数，e'_{qk} 是查询特征 z_q 经过线性投影得到的注意力分数，L_{qk} 为二维相对位置编码，r_{qk} 表示位置间的相对位置权值。这里相对位置编码应用时是与预过滤注意力模块中经全连接层输出的注意力打分 e'_{qk} 相加，再一同输入 Softmax 层输出最终的注意力分布 α_{qk}。

(2)映射函数

由于视觉 Transformer 输入序列较长，为避免相对位置编码带来参数和计算量的增加，我们引入一种多对一函数来构建相对距离到编码的映射方法，如式(6)所示。

$$F(x) = \begin{cases} sign(x) \times \min\left(\mu, \left[\lambda + \dfrac{\ln(|x|/\lambda)}{\ln(\nu/\lambda)}(\mu-\lambda)\right]\right), & |x| > \lambda \\ [x], & |x| \leqslant \lambda \end{cases} \tag{6}$$

其中，$[\cdot]$为取整运算，$sign(\cdot)$用来确定数的符号，通过调整 λ、μ、ν 控制分段点位置、输出范围以及对数部分的分配。函数将聚集在某一范围内有限的相对距离映射为同一编码，并通过相对距离的远近匹配不同的注意力，不仅减少参数和计算量，还避免了以往剪辑函数[15]中远距离位置上下文信息的丢失。

(3)二维相对位置的计算

视觉任务的输入通常是结构化图像数据，因此映射相对位置到编码时考虑其位置方向是很重要的，本研究提出了一种有向映射方法以定义相对权重 r_{qk} 进而计算相对位置，基于预过滤注意力模块的计算过程，利用 z_q 通过线性投影得到的偏移值进行映射计算，通过相对距离分配注意力，分别在水平和垂直方向上编码，最后将两者综合，此时相对位置编码由距离和方向一起决定，公式如下，其中 $G^x_{I^x(q,k)}$ 和 $G^y_{I^y(q,k)}$ 都是可学习标量，用以储存相对位置权重。

$$r_{qk} = G^x_{I^x(q,k)} + G^y_{I^y(q,k)} \tag{7}$$

$$I^x(q,k) = F(\Delta t^x_{hqk}) I^y(q,k) = F(\Delta t^y_{hqk}) \tag{8}$$

Δt^x_{hqk} 和 Δt^y_{hqk} 代表 z_q 分别在 x 轴和 y 轴上通过线性变换得到的偏移量，$F(\cdot)$ 即映射函数。

3)多尺度层面的预过滤注意力模块

现有的目标检测框架大多利用多尺度特征图预测不同尺度的目标物体，本研究中的模型同样可以延伸到多尺度层面，使用多尺度预过滤注意力模块。假设 $\{x^s\}_{s=1}^{S}$ 代表模型的多尺度输入特征，则相应的注意力计算为：

$$MSAttn(z_q, \hat{t}_q, x^s) = \sum_{h=1}^{H} W_h \left[\sum_{s=1}^{S}\sum_{k=1}^{K} \alpha_{hsqk} \cdot T^s W^V\right] \tag{9}$$

$$T^s = f^s(\varphi_s(\hat{t}_q) + \Delta t_{hsqk}) \tag{10}$$

其中，变量 s 代表输入特征层数，Δt_{hsqk} 和 α_{hsqk} 对应第 s 个特征层上的 Δt_{hqk} 和 α_{hqk}，计算方法仍与单尺度相同。第 s 个特征层上采样点的 value 值 T^s 仍由双线性插值 $f^s(\cdot)$ 计算得到。$\hat{t}_q$ 是 z_q 对应目标点归一化后的坐标，利用 $\varphi_s(\cdot)$ 函数将 $\hat{t}_q$ 映射到各个特征层中计算各特征层采样的位置。

此外，Transformer 检测模型输入延申到多个特征层时，不同特征层上特征点坐标可能相同，位置编码就容易混淆，因此我们在编码器输入中额外增加了一个可学习的位置嵌入 Layer Embedding(LE)以表明输入的多尺度性，使用时将各层的 Layer Embedding(LE)与对应公式计算的绝对位置嵌入 Position Embedding(PE)相加作为编码器的最终位置嵌入。

4)Transformer 内部架构

(1)编码器

与现有的 Transformer 检测模型不同，本研究提出的检测模型中，编码器的自注意力模块为预过滤注意力模块，此时编码器输入为 ResNet-50 下采样提取的多尺度特征图。query、key 和 value 均来自多尺度特征图输入，值得注意的是 query 结合的位置嵌入是 Position Embedding(PE)与 Layer Embedding(LE)之和。编码器中的目标点相当于输入多尺度特征点的归一化坐标，可以看作是 key 的位置信息，将该归一化坐标映射到所有特征层中以计算各层采样点位置。此外，前向反馈网络与原始 Transformer 相同，包含全连接层、激活函数、Dropout、残差连接以及层归一化，具体结构如图 4 所示。

(2)解码器

本研究提出的模型与原始 Transformer 的主要区别在于解码器结构中交叉注意力模块使用了预过滤注意力模

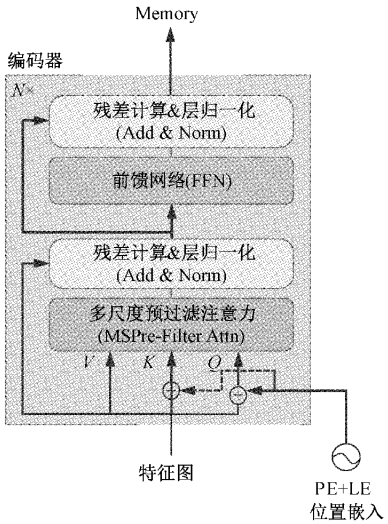

图 4　编码器内部结构

块,自注意力模块和前向反馈网络结构保持不变。与编码器不同,在解码器中目标点的二维归一化坐标由其预设的 Query Embedding 经线性投影获取,然后同样映射到各特征层上,这里的目标点也可视作 key 的位置信息。交叉注意力模块在改动后,其 Object query 依旧来自注意力层的输出,同时还要加上解码器最初的位置信息 Query Embedding,而 key 和 value 是由编码器输出的编码特征经线性变换获得,具体结构如图 5 所示。

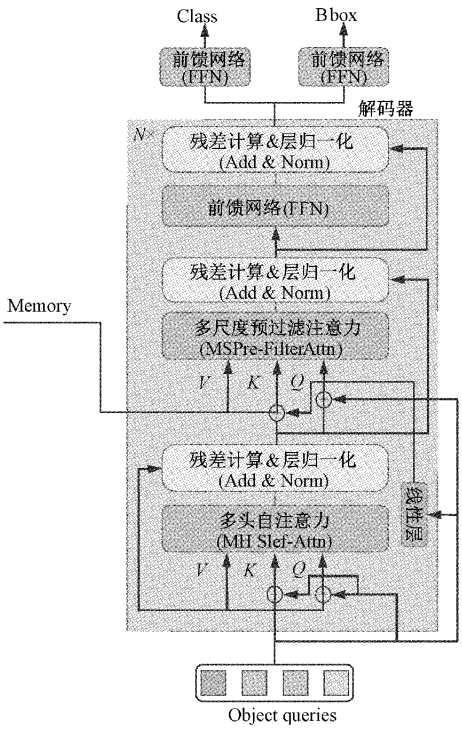

图 5　解码器内部结构

2.3　预测前馈网络

本研究的模型预测前馈网络由一个 3 层前馈神经网络和一个全连接层组成,分别用作回归边界框和识别物体种类。为了提高模型效率和性能,我们采用基于目标点的相对偏移值回归边界框。

3　实验及分析

3.1　实施细节

在 COCO 2017 检测数据集上对提出的模型进行性能验证,模型在训练集上训练,在验证集和测试集上进行性能评估,并报告 COCO 2017 验证数据集的标准平均精度(AP)等评价指标以分析模型的收敛速度和检测精度。

实验环境在 Unbuntu20.04 下配置,程序编写采用 Python3.7,并在 Pytorch 框架下搭建模型结构。模型采用 ImageNet 预训练的 ResNet-50 作为主干网络,在不使用 FPN 的前提下从中提取多尺度特征图作为输入。实验中设置预过滤注意力模块的相关参数 $H=8$ 和 $K=4$,用于获取目标点和采样偏移量的线性投影学习率的系数设为 0.1。采用类似偏差模式的有向相对位置编码方法将相对位置编码添加到 Transformer 所有预过滤注意力模块中,且编码在不同注意力头中共享。此外,相对位置映射函数的系数设置为 $\alpha:\beta:\gamma=1:2:8$。为了节省参数,规定模型编码器的参数在不同特征尺度下共享,其他参数预设值和训练方法仍沿用经典 Transformer 检测模型 DETR 中的设置,与 DETR 不同的是,用于边界框分类损失的损失权重修改为 2,将 Object query 的数量增加为 300。使用 Adam 优化器训练模型,初始学习率设为 2×10^{-4},权重衰减为 10^{-4},$\beta_1=0.9$,$\beta_2=0.999$,默认情况下模型在 4 个 NVIDIA Tesla V100 GPU 上从头开始训练 50 个 Epoch,batch_size 设为 1,学习率在第 40 个 Epoch 时衰减 0.1 倍。

3.2　基准比较

我们在 COCO 2017 数据集上进行了一系列实验以验证基于预过滤注意力模块的 Transformer 检测模型对检测性能的改善,实验分别选取了 7 种经典的基于 CNN 和 Transformer 的检测模型与文中提出的模型进行性能比较。由表 1 可以看出,模型 Faster R-CNN+FPN、DETR 和 DETR-DC5 都需要较长时间的训练才能达到收敛点,尤其是 DETR 检测小物体的效果非常不尽人意。值得注意的是,在与 Faster R-CNN+FPN、DETR 和 DETR-DC5 基线有着相似参数量的情况下,基于预过滤注意力模块的 Transformer 检测模型最多可以将模型训练时间缩短 10 倍,这大大加快了模型收敛速度,最终实现的性能也与 Faster R-CNN+FPN、DETR、DETR-DC5 相当甚至更好,尤其是在小物体检测上最高提升了 6.3 APs,增强了模型的定位能力和检测精度。与另外 3 种先进的 Transformer 检测模型进行比较,在训练 Epoch 和参数量相当的情况下,基于预过滤注意力模块的 Transformer 检测模型也在不同程度上提高了模型检测的精确度,并在小物体检测上达到了 26.8 APs 的最优结果,这都显示出基于预过滤注

表 1　COCO 数据集上各基线性能

方法	Epochs	Params/M	FLOPs/G	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN+FPN	109	42	180	42.0	62.1	45.5	26.6	45.4	53.4
DETR	500	41	86	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	500	41	187	43.3	63.1	45.9	22.5	47.3	61.1
Deformable DETR	50	40	173	43.8	62.6	47.7	26.4	47.1	58.0
Conditional DETR	50	44	90	40.9	61.8	43.3	20.8	44.6	59.2
TSP-FCOS	36	—	189	43.1	62.3	47.0	26.6	46.8	55.9
本文方法	**50**	**40**	**175**	**44.1**	**63.4**	**47.9**	**26.8**	**47.7**	**59.3**

意力模块的 Transformer 检测模型的可行性，通过改变注意力模块的计算方式，融入有向相对位置编码，有效解决了经典 Transformer 检测模型的局限性，提高模型收敛速度和小物体检测的精度，同时也证明了相对位置编码在目标检测中的有效性。

3.3　消融实验

1)注意力模块架构

设定模块输入均为多尺度特征，我们针对模型中注意力模块的架构进行了相应的消融实验。由表 2 可知，使用多尺度层面下的预过滤注意力模块可以在不利用 FPN 的情况下实现不同尺度特征图之间的信息融合，不仅加快模型的收敛速度，同时还带来 1.6 AP 的提高，特别是在小物体检测方面有效提升了 2.4 APs，证明了预过滤注意力模块应用到多尺度层面的有效性。另外，由于基于预过滤注意力模块的 Transformer 检测模型中的注意力模块已具备多尺度特征间信息交换的能力，所以我们强调不再需要 FPN 来组合多尺度特征，这一点也可以通过消融实验看出，即使再加入加权双向特征金字塔网络(bidirectional feature pyramid network，BiFPN)[20]也没有对模型的性能产生明显的改善。

表 2　注意力模块架构消融实验

MS attention	FPN	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	w/o	42.5	62.2	46.2	24.4	45.7	57.6
√	w/o	**44.1**	**63.4**	**47.9**	**26.8**	**47.7**	**59.3**
√	BiFPN	44.2	63.3	47.9	25.9	48.1	59.0

2)相对位置信息

针对注意力模块中添加的有向相对位置编码，我们也对编码计算中的几个步骤进行了相应的消融实验，以显示有向相对位置编码的效果。

(1)相对位置编码

传统只含绝对位置编码的检测模型仍有良好的性能，但在这基础上通过添加有向相对位置编码，模型获得更丰富的位置信息，性能可以在一定程度上提升。表 3 中显示在 50 个训练 Epoch 下，添加了相对位置编码的方法将只含有绝对位置编码的 Transformer 检测模型性能持续提高了 0.7 AP，在小物体检测上的精度也达到了 26.8 APs。

表 3　相对位置编码消融实验

Abs Pos.	Rel Pos.	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
sinusoid	none	50	43.4	62.5	47.2	26.1	46.8	58.1
sinusoid	bias	**50**	**44.1**	**63.4**	**47.9**	**26.8**	**47.7**	**59.3**
sinusoid	bias clip	50	43.7	62.8	47.5	27.2	46.9	58.6

(2)映射函数

实验中对比了两种映射函数的效果，如表 3 所示，在相同的 Epoch 下看到，剪辑函数[15]的效果并不如 $F(\cdot)$ 映射函数，$F(\cdot)$ 映射函数的使用可以进一步提升模型 0.4 AP 的检测精度。分析其主要原因是当输入序列较短时，两个映射函数的作用相当，但由于目标检测任务的输入序列相对较长，这时 $F(\cdot)$ 映射函数的优势就逐渐体现出来，它将不同的注意力分配到距离相对较远的位置，避免远距离位置信息的损失，有利于提高模型检测的精确度。

3.4　可视化实验

本研究对模型编码器、解码器末层的采样点和注意力权值进行了可视化实验，以便清晰的展示和理解预过滤注意力模块。我们从不同尺度的特征图中提取采样点和注

意力权重,并整合到一张图中进行观察分析,如图 6、7 所示。模型中选取的采样点在图 6 中展示为实心点,实心点的颜色则表示对应采样点注意力权重的大小,红色代表权重值高,蓝色代表权重值低。图 7 中十字符号表示模块计算中的目标点。由图 6、7 分析可知,物体在基于预过滤注意力模块的 Transformer 检测模型的编码器中已被分离,而在解码器中,不同于 DETR 中只关注极值点,本研究提出模型更加关注图片完整的前景实例,这表明我们的模型不仅需要极值点,还同时利用内部的点来进行物体的检测。

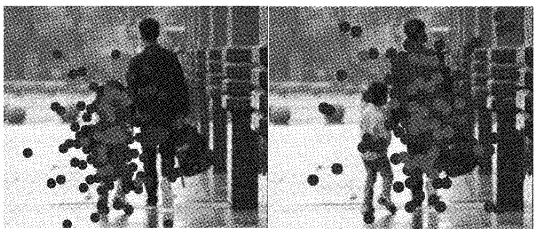

图 6　编码器预过滤注意力模块的可视化

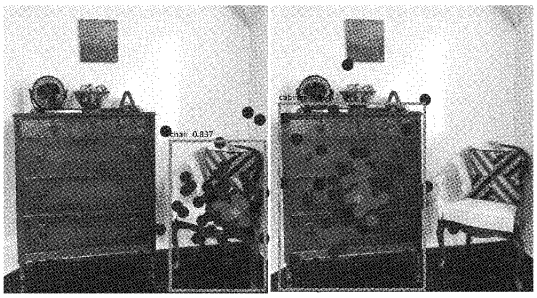

图 7　解码器预过滤注意力模块的可视化

此外,实验还表明基于预过滤注意力模块的 Transformer 检测模型使用的多尺度预过滤注意力模块能根据前景对象的不同尺度和形状调整其采样点和注意力权重。

4　结　　论

本研究分析目前基于 Transformer 的检测模型的局限及其原因,提出了一种基于预过滤注意力模块的 Transformer 检测模型,其注意力模块的核心为稀疏定位思想和相对位置信息,它有效的解决了模型由于全局关注导致的收敛缓慢,实现更好的小物体检测性能。在 COCO 2017 数据集上的实验也取得了令人满意的结果,证明了我们的方法在目标检测任务中的有效性,同时显示了相对位置编码在目标检测领域的有效性。未来的工作中我们尝试进一步将其应用到其他场景,希望可以促进对注意力机制和 Transformer 原理的深层次研究,特别是对全局注意力的利用和改进。

参考文献

[1]　周晓彦,王珂,李凌燕. 基于深度学习的目标检测算法综述[J]. 电子测量技术,2017,40(11):89-93.

[2]　REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[3]　LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. IEEE Trans Pattern Anal Mach Intell, 2020, 42(2): 318-327.

[4]　TIAN Z, SHEN C, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]. 2019 IEEE International Conference on Computer Vision (ICCV 2019), 2019: 9626-9635, DOI: 10.1109/ICCV.2019.00972.

[5]　徐印赟,江明,李云飞,等. 基于改进 YOLO 及 NMS 的水果目标检测[J]. 电子测量与仪器学报,2022,36(4):114-123.

[6]　VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. International Conference on Neural Information Processing Systems (NIPS 2017), 2017: 6000-6010, DOI: 10.18653/v1/2020.emnlp-main.317.

[7]　LIN T, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), 2017: 936-944, DOI: 10.1109/CVPR.2017.106.

[8]　LIN T, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]. European Conference on Computer Vision(ECCV 2014), 2014: 740-755, DOI: 10.1007/978-3-319-10602-1_48.

[9]　解尧婷,张丕状. 基于改进的 YOLOv4 输电线路小目标检测[J]. 国外电子测量技术,2021,40(2):47-51.

[10]　CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. European Conference on Computer Vision (ECCV 2020), 2020: 213-229, DOI: 10.1007/978-3-030-58452-8_13.

[11]　DAI Z, CAI B, LIN Y, et al. Up-detr: Unsupervised pre-training for object detection with transformers[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2020), 2020: 1601-1610, DOI: 10.1109/CVPR46437.2021.00165.

[12]　DAI X Y, CHEN Y P, YANG J W, et al. Dynamic DETR: End-to-end object detection with dynamic attention[C]. IEEE International Conference on Computer Vision (ICCV 2021), 2021: 2968-2977, DOI: 10.1109/ICCV48922.2021.00298.

[13] LI F, ZHANG H, LIU S, et al. DN-DETR: Accelerate DETR training by introducing query DeNoising[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2022), 2022: 13619-13627, DOI: 10.48550/arXiv.2203.01305.

[14] 杨丽,吴雨茜,王俊丽,等.循环神经网络研究综述[J].计算机应用,2018,38(S2):1-6,26.

[15] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[C]. North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL 2018), 2018: 464-468, DOI: 10.18653/v1/N18-2074.

[16] HUANG Z, LIANG D, XU P, et al. Improve transformer models with better relative position embeddings[C]. In Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing(EMNLP 2020), 2020: 3327-3335, DOI: 10.18653/v1/2020.findings-emnlp.298.

[17] DAI Z, YANG Z, YANG Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[C]. Association for Computational Linguistics (ACL 2019), 2019: 2978-2988, DOI: 10.18653/v1/P19-1285.

[18] RAMACHANDRAN P, PARMAR N, VASWANI A, et al. Stand-alone self-attention in vision models[C]. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019: 68-80, Corpus ID: 189897750.

[19] HE K, ZHANG X, REN S, SUN J. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2016), 2016: 770-778, DOI: 10.1109/CVPR.2016.90.

[20] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020), 2020: 10781-10790, DOI: 10.1109/cvpr42600.2020.01079.

作者简介

王琪,研究生,主要研究方向为人工智能、图像处理等。
E-mail:1554434867@qq.com

赵文仓(通信作者),博士,教授,主要研究方向为认知智能、图像处理、智能科学与技术等。
E-mail:17864283900@163.com