多粒度空间注意力与空间先验监督的DETR

廖峻霜, 谭钦红

引用本文

廖峻霜, 谭钦红. 多粒度空间注意力与空间先验监督的DETR[J]. 计算机科学, 2024, 51(6): 239-246.

LIAO Junshuang, TAN Qinhong. DETR with Multi-granularity Spatial Attention and Spatial Prior Supervision [J]. Computer Science, 2024, 51(6): 239-246.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300218

多粒度空间注意力与空间先验监督的 DETR

廖峻霜　谭钦红
重庆邮电大学通信与信息工程学院　重庆 400065
(s210131121@stu.cqupt.edu.cn)

摘　要　近年来,Transformer 在视觉领域的表现卓越,由于其优秀的全局建模能力以及可媲美 CNN 的性能表现受到了广泛关注。DETR(Detection Transformer)是在其基础上研究的首个在目标检测任务上采用 Transformer 架构的端到端网络,但是其全局范围内的等价建模以及目标查询键的无差别性导致其训练收敛缓慢,且性能表现欠佳。针对上述问题,利用多粒度的注意力机制替换 DETR 的 encoder 中的自注意力以及 decoder 中的交叉注意力,在距离近的 token 之间使用细粒度,在距离远的 token 之间使用粗粒度,增强其建模能力;并在 decoder 中的交叉注意力中引入空间先验限制对网络训练进行监督,使其训练收敛速度得以加快。实验结果表明,在引入多粒度的注意力机制和空间先验监督后,相较于未改进的 DETR,所提改进模型在 PASCAL VOC2012 数据集上的识别准确度提升了 16%,收敛速度快了 2 倍。

关键词:多粒度空间注意力;空间先验监督;目标检测;视觉 Transformer;编解码架构
中图分类号　TP391.4

DETR with Multi-granularity Spatial Attention and Spatial Prior Supervision

LIAO Junshuang and TAN Qinhong
School of Communication and Information Engineering,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　The Transformer has shown remarkable performance in the field of computer vision in recent years,and has gained widespread attention due to its excellent global modeling capability and competitive performance compared to convolutional neural networks(CNNs). Detection Transformer(DETR) is the first end-to-end network that adopts the Transformer architecture for object detection tasks,but it suffers from slow convergence during training and suboptimal performance due to its equivalent modeling across the global scope and indistinguishability of object query keys. To address these issues,we propose replacing the self-attention in the encoder and the cross-attention in the decoder of DETR with a multi-granularity attention mechanism,using fine-grained attention for tokens that are close in distance and coarse-grained attention for tokens that are far apart,to enhance its modeling capability. We also introduce spatial prior constraints in the cross-attention of the decoder to supervise the network training,which accelerates the convergence speed. Experimental results show that the proposed improved model,after incorporating the multi-granularity attention mechanism and spatial prior supervision,achieves a 16% improvement in recognition accuracy on the PASCAL VOC2012 dataset compared to the unmodified DETR,with a doubled convergence speed.

Keywords　Multi-granularity spatial attention,Spatial prior supervision,Object detection,Vision Transformer,Encoder-Decoder architecture

近年来深度学习发展迅猛,计算机视觉领域、自然语言处理等领域也因此取得了长足的发展。在计算机视觉领域,由于卷积操作中的卷积核的参数共享以及归纳偏置特点与图片的局部特征特点相吻合,在 Transformer[1] 应用于视觉领域之前,所有的计算机视觉任务网络以卷积神经网络(CNN)架构为主,例如一阶段目标检测器 YOLO[2] 系列、Retina-Net[3] 等,二阶段的 Faster RCNN[4]、后续基于无锚框的算法 CenterNet[5] 以及 CornerNet[6],它们都是以 CNN 作为主干网络搭建的深度神经网络。采用 CNN 架构在各种视觉任务中都取得了前所未有的成功,包括图片分类、目标检测、实例分割等计算机视觉任务。在基于 CNN 的网络架构下,很多针对目标检测任务的优化方法被提出,例如对小目标检测性能有较大提升的多尺度特征预测 FPN[7] 网络,针对 FPN 的单向自上而下的路径导致其特征融合不完全改进的 PANet[8],对卷积特征的通道和空间上引入注意力机制的 CBAM[9]。CNN 在多年的发展中逐渐成熟。

近几年,Transformer 架构在 NLP 领域中获得极大发展,其在 NLP 领域中的作用、地位与 CNN 在图像中扮演的角色类似,因此越来越多的学者开始探索其在视觉领域的应用。在图像分类任务上,VIT[10] 首次提出将原始的 Transformer 架构直接应用于视觉网络模型中,其将输入大小为 224×224 的图片划分为 16×16 固定大小的 14×14 个 patch 并加上相应的绝对位置编码形成输入,送入未经调整的原始 Transformer 的 encoder 中并加入 class token(CLS)进行训练,最终

通信作者:谭钦红(tanqh@cqupt.edu.cn)

将 class token 送入分类器中进行分类，在大规模的数据集且长时间的训练下，其分类性能达到了与以 CNN 为主干的 SOTA 网络相当的性能。针对 ViT 中需要庞大的数据集来进行长时间训练的问题，DeiT[11] 在数据集以及数据增强方面使用了基于知识蒸馏的教师-学生模型，从而在较小数据集上依然取得了较好的效果。PvT[12] 和 CvT[13] 分别利用映射和卷积操作，在 ViT 的基础上将各个 patch 进行空间尺度上的放缩以及通道数的改变，形成类似于传统 CNN 架构中的多尺度架构，从而提高网络性能。Li 等[14] 提出了另一种多尺度的架构 tokens-to-token，他们认为 ViT 的很多 token 之间其实包含很多的相互信息，利用多个 Transformer 块将多个 tokens 逐步提炼成一个 token，从而提高网络的性能。

在目标检测任务上，DETR[15] 是研究者提出的第一个端到端目标检测网络，与用于图片分类的 ViT 不同的是，它将输入图片先经过 CNN 主干网络（如 ResNet50[16]）进行特征提取，然后将提取的特征进行嵌入映射后送入完整的 Transformer 编码器解码器架构中进行注意力计算。这种算法将目标检测看作是一个目标框的集合预测问题，在输出头中采用匈牙利算法对预测框进行二分匹配，因此消除了常规 CNN 架构中的繁琐人工步骤，即网络进行前的锚框设置以及用于处理检测重复框的非极大值抑制（NMS），真正做到了端到端的目标检测训练。但是，DETR 由于仍然采用原始的 Transformer 编解码架构，利用的是 Transformer 的全局等价建模能力，所有的 token 都需要做注意力计算，并且在 decoder 中 object-query 需要同时进行空间与内容上的特征学习，因此其训练收敛速度缓慢。鉴于此，本文提出了基于多粒度的注意力机制以及空间先验监督的 DETR，以期在改善检测性能的同时加快收敛速度。本文的主要贡献如下：

1）针对识别精度、准确度不高的问题，本文将 Transformer的 encoder 中的自注意力模块以及 decoder 中的交叉注意力模块替换为多粒度的注意力机制，进一步提升模型的检测能力。

2）针对训练收敛速度慢的问题，提出在 decoder 中的交叉注意力模块中引入空间先验限制对网络训练进行监督训练，使其 decoder 中的 object-query 优先收敛于其先验空间，从而加速训练收敛速度。

1 相关理论介绍

1.1 DETR 架构

标准的 detection Transformer 架构由三大部分组成：编码器（Encoder）和解码器（Decoder）以及预测头（Prediction）。其中编码器由 L 个重复堆叠的多头自注意力（Multi-head Self Attention，MSA）模块以及前向传递网络（FFN）组成；解码器由多头自注意力模块（MSA）、交叉多头自注意力模块（Multi-head Cross-Self Attention）和前向传递网络组成；预测头由两个简单的带有 ReLu 激活函数的 MLP 组成。数据前向流程如图 1 所示，输入一个 Batch 的数据集图片（$B\times H_0\times W_0\times C_0$），由 ResNet50 或者其他 CNN 主干网络进行特征提取，提取的特征向量表达为 $B\times H_1\times W_1\times C_1$。对于 ResNet50 而言，$H_1=H_0/32$，$W_1=W_0/32$，$C_1=2\,048$），提取后的特征再通过 1×1 的卷积（对应的参数为 in channel$=C_1$，out channel$=D$，kernel size$=1$，stride$=1$）进行通道数改变，再映射到 Transformer 的输入嵌入空间 $\boldsymbol{X}(B\times D\times(H_1\times W_1)$，$D$ 为嵌入的特征维度），然后送入 Transformer 中的编解码器，最终通过 FFN 预测头得到预测输出。

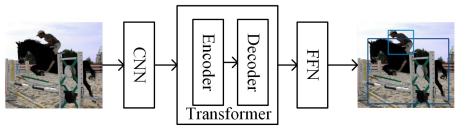


图 1 DETR 前向传递过程

Fig. 1 Forward process of DETR

1.1.1 编码器

编码器的主要作用是提取前级 CNN 网络提取的特征图的内容信息，计算每个特征之间的相关性，由于 Transformer 的输入需要具有序列信息，因此还要对卷积提取的输出进行编码器输入映射形成 N 个输入特征，$E^i\in R^{B\times D\times H\times W}$，$i\in(1,N)$。又因为 Transformer 架构本身具有分布不变性，对输入的每个特征的序列位置并不敏感，为了更好地区分每个特征，需要加上其对应的绝对位置编码 $E_{\text{pos}}\in R^{B\times D\times(H\times W)}$，加上位置编码后形成编码器输入的初始向量 $\boldsymbol{Z}_0$，然后在 L 个 Encoder Layer 中先后经过多头自注意力模块（MSA）以及前向传递网络（FFN）形成编码器输出（$Z_L\in R^{B\times D\times(H\times W)}$），LN 为增强网络训练稳定性的 LayerNorm 操作，$\ell\in(1,\cdots,L)$，DETR 中使用的编码器架构如图 2 所示，编码器计算式如下：

$$\boldsymbol{Z}_0=(E^1,E^2,\cdots,E^N)+E_{\text{pos}} \tag{1}$$

$$Z_\ell'=LN(MSA(Z_{\ell-1})+Z_{\ell-1}) \tag{2}$$

$$Z_\ell=LN(FFN(Z_\ell')+Z_\ell') \tag{3}$$

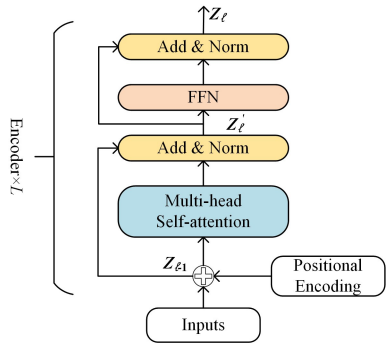


图 2 编码器架构

Fig. 2 Structure of encoder

1.1.2 解码器

解码器中的多头自注意力计算表达式与编码器中的表达式一致，目的是消除解码器中查询对于目标的重复预测，解码器架构如图 3 所示。预定义的特征维度为 D 的 M 个 Object Query（$Q\in R^{M\times D}$，$Q^i\in R^D$，$i\in(1,M)$），通过和 Query Embedding（$QE\in R^{M\times D}$）相加后形成解码器的多头自注意力模块（MSA）的输入 X_0，经过 MSA 输出后再与编码器的嵌入输出 Z_L 一起输入到多头交叉注意力模块（M-CSA）中进行计算得到 X_L，最后通过一个分类头和边框回归头进行输出，表达式如下：

$$X_0=(Q^1,Q^2,\cdots,Q^M)+QE \tag{4}$$

$$X_\ell''=LN(MSA(X_{\ell-1})+X_{\ell-1}),\ell\in(1,\cdots,L) \tag{5}$$

$$X_\ell'=LN(MCSA(X_\ell'',Z_L)+X_\ell''),\ell\in(1,\cdots,L) \tag{6}$$

$$X_\ell=LN(FFN(X_\ell')+X_\ell'),\ell\in(1,\cdots,L) \tag{7}$$

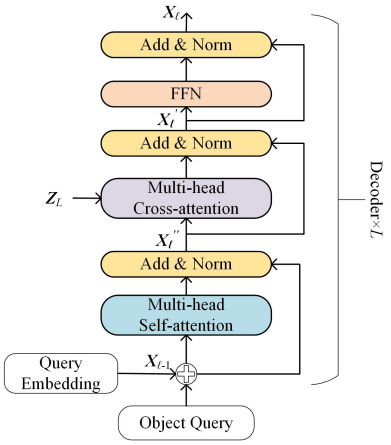


图 3　解码器架构

Fig. 3　Structure of decoder

1.1.3　预测输出头

在解码器完成计算输出 X_L 后，分别使用一个类别预测头和一个边界框位置回归头对其进行输出。两个预测头都由简单的单层 MLP 组成，Cls 代表预测输出类别数，Box 代表网络预测的输出目标框，表达式如下：

$$Cls=\mathrm{Sigmoid}(MLP(X_L)) \tag{8}$$

$$Box=MLP(X_L) \tag{9}$$

预测输出的 Cls 通过一个 Sigmoid 函数进行预测类别输出，Box 由物体的中心点(x,y)以及宽高(w,h)组成，范围为未归一化的$[0,1]$，然后经过反归一化进行坐标映射，映射回原始图像相对应的坐标即完成预测。

1.2　多头自注意力机制

Transformer 中使用注意力机制的目的是计算矩阵元素之间的相关性，从而筛选出与其最相关的特征元素，每个 token 或者 patch 即为矩阵元素。通常，输入矩阵通过 3 个不同的映射矩阵 $\boldsymbol{W}^Q\in R^{D\times d_q}$，$\boldsymbol{W}^K\in R^{D\times d_k}$，$\boldsymbol{W}^V\in R^{D\times d_v}$ 形成 $\boldsymbol{Q}\in R^{B\times D\times d_q}$，$\boldsymbol{K}\in R^{B\times D\times d_k}$，$\boldsymbol{V}\in R^{B\times D\times d_v}$，$D$ 为特征嵌入的维度，d_q，d_k，d_v 分别为对应 $\boldsymbol{QKV}$ 矩阵的映射维度，然后经过注意力打分函数（缩放点积打分函数）进行注意力计算，最后将注意力得分即相关性与 $\boldsymbol{V}$ 进行计算，公式如下：

$$\boldsymbol{Q}=X\boldsymbol{W}^Q,\boldsymbol{K}=X\boldsymbol{W}^K,\boldsymbol{V}=X\boldsymbol{W}^V$$

$$Z=\frac{\mathrm{Softmax}(\boldsymbol{Q}\times\boldsymbol{K}^{\mathrm{T}})}{\sqrt{d_k}}\boldsymbol{V} \tag{10}$$

图 4 展示了单头注意力的计算过程，但采用一个头进行自注意力进行计算限制了模型的表达能力，其整体的映射空间受限。多头自注意力机制是为了扩展整个模型的表征空间，增强其表达能力而设计的，具体做法是将参与计算的 $\boldsymbol{QKV}$ 向量在最后的维度进行 h 个头分割，然后每个头分别进行自注意力计算，再将输出的结果进行 concatenate，形成最后的输出。将式(10)修改为式(11)。

$$Z=Cat\left(\frac{\mathrm{Softmax}(Q_i\times K_i^{\mathrm{T}})}{\sqrt{d_{ki}}}V_i\right),i\in(1,\cdots,h) \tag{11}$$

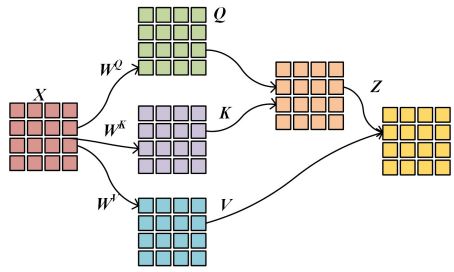


图 4　注意力机制计算

Fig. 4　Attention mechanism computing

2　主要问题及相关工作

DETR 中设计了一个新型的基于 Transformer 架构的目标检测架构，虽然摒弃了相对于 CNN 架构网络而言繁琐的后处理以及锚框设计，但是其收敛速度仍然很慢，对于小目标的检测性能较差。其原因主要在于：1）解码器中的 object query 需要同时学习编码器输出中的内容信息以及位置信息，在未经长时间的训练下，很难从中学习到足够的空间信息；2）对于小目标而言，Transformer 中全局平均注意力机制覆盖的图像面积区域与自注意力计算的感受野不成正比，导致太多冗余无用信息被提取出来。

解码器中的交叉注意力模块的主要作用是进一步提取前级编码器的输出嵌入。最开始其将注意力全部平均到整个特征图上，要学习到最后感兴趣的物体稀疏的位置需要长时间的训练。图 5 给出了 DETR 最终学习到的部分 object query 所表达的位置和内容信息，可以发现，其注意力分布基本上是学习到了数据集中检测物体的分布，它们各自关注自己的那部分稀疏空间。由此可见，object query 所起到的作用与锚框在 CNN 网络架构中的作用类似，都是对空间的一种限制。和本文研究相关的其他文章也做了类似的改进工作。Deformable DETR[17] 针对网络架构中的注意力计算机制提出了一个核心改进点：让当前 token 或者 patch 只关注受影响较大的几个稀疏点，而不是像普通注意力机制那样关注所有的空间点，因此，其利用基于可变形卷积（Deformable Convolution）的多头可变形注意力模块替代编码器的多头自注意力模块以及解码器中的交叉注意力模块，并利用多尺度的卷积特征来学习与其相关度最高的几个稀疏点，进一步提高小目标的检测性能，并且加快其训练过程。可分离卷积模块的表达式如下：

$$Attn=\sum_{h=1}^{H}W_h\left[\sum_{k=1}^{K}A_{hqk}\cdot W_h'x(p_q+\Delta p_{hqk})\right] \tag{12}$$

其中，H 表示注意力头的个数，K 表示可分离卷积关注的稀疏点，p_q 与 Δp_{hqk} 分别代表卷积网络提取后特征图上的参考点 $p(x,y)$ 及对应的偏移 $\Delta p_{hqk}\in R^{(K\times 2)}$，$\boldsymbol{W}_h'$ 和 $\boldsymbol{W}_h$ 分别代表式(10)中的 $\boldsymbol{W}^V$ 映射矩阵以及最后的相关性映射矩阵，A_{hqk} 是通过注意力打分函数计算出的注意力权重，其计算式同式(10)。

SMCA[18] 认为每层的解码器嵌入输出携带着物体的位置信息和类别信息（最终网络的预测输出头也是基于此来进行预测的），所以使用了另外一个线性网络来提前获得每个 object query 所对应的未归一化位置(c_w,c_h)以及对应尺度

(s_w, s_h)，并将其进行类高斯特征图映射后加入 MSA 模块中。其中 β 是可调的超参，用于调整整个高斯图的方差，也就是使用空间调制协同注意力模块替换交叉注意力模块，进而加速网络训练。使用类高斯特征图映射对注意力进行修改，修改后的表达式如式(13)和式(14)所示。

$$G(i,j)=\exp\left(-\frac{(i-c_w)^2}{\beta s_w^2}-\frac{(j-c_h^2)}{\beta s_h^2}\right) \tag{13}$$

$$Z=Cat\left(\frac{\mathrm{Softmax}(Q_i\times K_i^T+\log G)}{\sqrt{d_{ki}}}V_i\right) \tag{14}$$

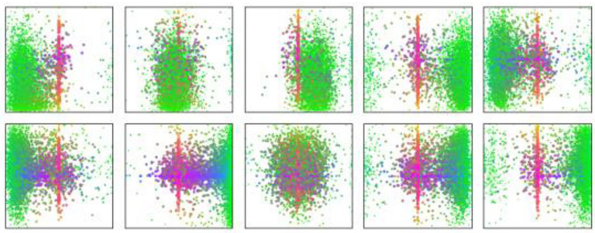


图 5 DETR 中部分 object query 的注意力分布

Fig. 5 Attention distribution of object query in DETR

3 本文方法

3.1 基于空间先验的解码器

和 SMCA 中的观点类似，若能将 object query 的表达空间限制在一个较小的范围内，使其从编码器的输出嵌入中学习到更关注于内容上的信息，则能加快整个网络的收敛速度。和其不同的是，SMCA 是将 object query 学习到的位置信息和尺度信息进行高斯权重映射，映射到与解码器嵌入的相同空间内并采用对数缩放对其进行元素相加。本文采用的是一种更直接的方法：直接采用 MLP 对目标查询(Object Query，OQ)进行参考坐标点学习，生成坐标映射，然后将该坐标与正弦位置嵌入(Positional Embedding，PE)相加并映射到对应的嵌入空间形成参考点映射(Reference Point，RP)，然后与解码器自注意力模块的输出进行 concate，形成 MCSA 的查询输入(Q_2)，以此完成基于空间先验的监督训练。加入空间先验后的结构如图 6 所示。

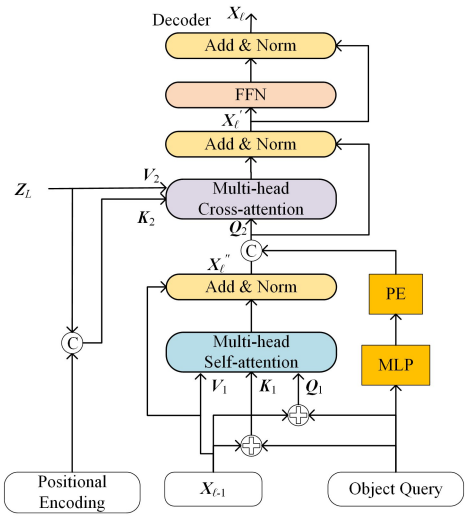


图 6 空间先验的表征

Fig. 6 Presentation of spatial prior

图 7 展示了未加入空间监督下的随机初始化注意力权重的感受野(左)以及加入了空间先验后对模型的监督训练(右)。在未加入空间先验监督时，原来的 object query 注意力分布是随机初始化的，图中显示的是一种特殊情况，且所有的 object query 未全部画出，在这种情况下模型需要较长的迭代时间才能使对应的查询学习到对应的稀疏注意力。而在加入了先验监督后，整个 object query 的收敛集合在训练过程中逐渐收敛于限定空间，极大地加快了模型的训练过程。

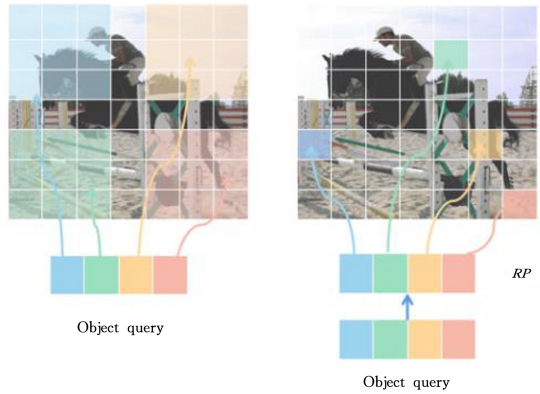


图 7 加入空间先验前后的注意力权重分布

Fig. 7 Attention weight distribution before and after adding spatial prior

解码器的 object query 初始化为 X_0，加入空间先验知识并更改进行注意力计算的相应矩阵，表达式如下：

$$X_0=Query\ Embedding \tag{15}$$

$$X_\ell''=LN(MSA(Q_1,K_1,V_1)+X_{\ell-1}) \tag{16}$$

$$RP=PE(MLP(OQ)) \tag{17}$$

$$Q_2=Concate(X_\ell'',RP) \tag{18}$$

$$X_\ell'=LN(MCSA(Q_2,K_2,V_2)+Q_2) \tag{19}$$

$$X_\ell=LN(FFN(X_\ell')+X_\ell') \tag{20}$$

式(17)中的 PE 算子采用绝对正弦位置编码。对每个位置上的元素而言，在每个不同的奇偶位置上编码是不同的，D 为特征嵌入的维度，偶数位置$(2i)$使用正弦函数，奇数位置$(2i+1)$使用余弦函数进行编码 $i\in[0,D/2]$。其编码计算式如下：

$$PE(2i)=\sin\left(\frac{1}{10\,000^{2i/D}}\right) \tag{21}$$

$$PE(2i+1)=\cos\left(\frac{1}{10\,000^{2i/D}}\right) \tag{22}$$

3.2 多粒度空间注意力

DETR 编码器中多头自注意力模块的作用主要是更好地计算由 CNN 网络提取并进行映射后的输入嵌入各个元素之间的相关性。采取的方式是对特征图上全局像素进行一一对应的无差别的计算机制，这种方式相对于输入特征图的宽高而言复杂度是平方级的，对于视觉领域而言，这是难以接受的，尤其对于小目标的检测而言。小目标的语义信息只能从低层特征图获得，而低层特征图的分辨率又较大，如果将高分辨率特征图用于常规的注意力计算，那么花费的开销将是巨大的。另一方面，图像中的目标或者特征具有局部性这一先验知识，我们很容易知道与某个特征图上的某个像素点相关

性最大的一些区域或者一些集合点应该分布在其周围，对于远处的一些像素点或者区域，应该施以较小的甚至可以忽略的注意力权重。Yang 等[19]在 ViT 的基础上，将多个 patch 之间的注意力计算更改为多粒度的注意力计算方式，根据距离的远近选择不同的注意力权重进行计算，其在相同的覆盖率下相比常规注意力使用了更少的 tokens，从而提高了图片分类的性能。受此启发，本文提出在原始的注意力计算机制的基础上，结合多粒度的空间注意力机制，对距离近的元素在计算时赋予更大的权重，对距离稍远一些的像素点赋予较小的权重，对距离很远的像素点赋予最小的注意力权重，权重集合 $\alpha_i \in \alpha, i \in [0,1,2]$，$\alpha=[0.5,0.3,0.2]$ 作为超参数可以对其进行调整。图 8 展示了根据当前像素 (x_0,y_0) 与其他像素 (x,y) 之间的距离选择不同的注意力权重 $Atten(x_0,y_0)$ 的多粒度注意力计算机制，分为 3 种粒度来选择注意力。A_1 区域内的像素点采用第一种粒度的注意力权重，A_2 以及 A_3 区域分别采用另外两种粒度的注意力。(h,w) 为输入特征图的宽高。

$$D=\min(h,w) \tag{23}$$

$$dist(P_1,P_0)=\begin{cases}0, & \text{if } L(P_1,P_0)<D/3\\ 1, & \text{elif } L(P_1,P_0)<2*D/3\\ 2, & \text{else}\end{cases} \tag{24}$$

$$Atten(x_0,y_0)=\alpha_i, i=dist((x,y),(x_0,y_0)) \tag{25}$$

具体地，首先找到当前特征图的宽高中的较小值作为基准值 D，然后在后续的注意力计算模块当中，根据要计算注意力的点 P_1 的位置与当前点 P_0 的距离 $L(P_1,P_0)$ 不同可以选择出 3 个不同的区域，进而决定注意力权重的取值 $dist(P_1,P_0)$，一个是 A_1，即 $L(P_1,P_0)$ 距离小于 $1/3*D$ 的正方形区域内，此时 $dist(P_1,P_0)$ 取值为 0。在注意力权重超参集合 α 中取 $Atten(x_0,y_0)=\alpha[0]$，采用较大的粒度来进行注意力机制计算，距离在 $1/3*D$ 到 $2/3*D$ 之间的矩形采用中间值粒度 $Atten(x_0,y_0)=\alpha[1]$ 进行计算；距离在 $2/3*D$ 以外的矩形采用最小的粒度 $Atten(x_0,y_0)=\alpha[2]$ 进行计算。显然，这种计算方式更加符合图像本身的性质，即越远的像素之间相关性越小，越近的像素之间相关性越大。这对于目标检测的性能提升更为重要。

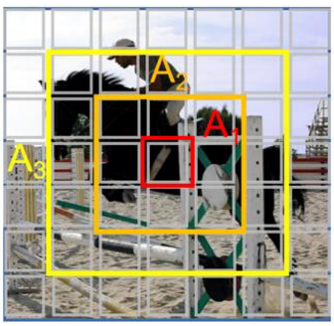


图 8　多粒度注意力机制

Fig. 8　Attention mechanism of multi-granularity

3.3　整体改进模型架构

在解码器部分添加空间先验监督以及将编码器的 MSA 和解码器的 MCSA 替换为多粒度空间注意力模块（Multi-Granularity Self Attention，MGSA）后的整体网络架构如图 9 所示。改进部分主要是将原来的自注意力计算模块替换成所提出的多粒度空间注意力模块，并在解码器的自注意力模块加入了使用线性映射和位置编码的空间先验监督机制。

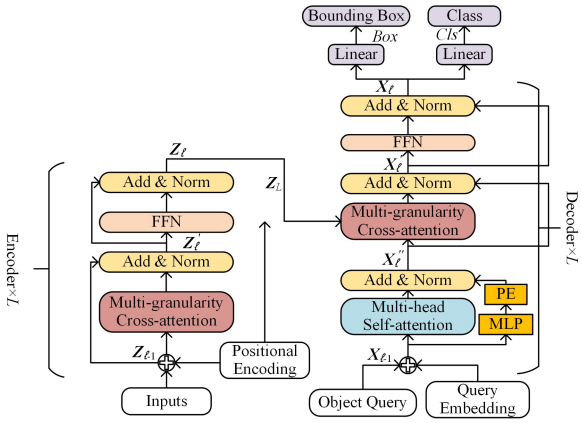


图 9　改进的网络模型架构

Fig. 9　Advanced network architecture

4　实验结果与分析

为验证提出的基于多粒度空间注意力机制以及空间先验监督训练的有效性，本实验采用公开数据集——PASCAL VOC2012 目标检测数据集作为实验数据集，该数据集共有 4 个大类，20 个小类，验证集与训练集总计有逾 11 000 张图片，其中训练集包含为 5 717 张图片，验证集包含 5 823 张图片，总的标注实例超 30 000 个。

4.1　实验设置及评价指标

4.1.1　实验设置

整个实验环境及设置如下：实验主机系统为 Ubuntu 20.04.1 LTS 系统，python 版本为 3.8，深度学习框架采用基于 python 的 Pytorch 1.13.1 版本，显卡采用 NVIDIA 3090，cuda 版本使用 11.7。数据增强手段采用常规的数据增强，即随机裁剪、翻转、缩放等，以提高模型的泛化能力。对照模型选用 DETR-DC5，其采用的特征提取 CNN 网络为 ResNet50。学习率设置为 0.0001，学习率衰减 10%，batch size 设置为 8，采用 AdamW 优化器，在倒数 20 个 epoch 开始对学习率进行调整，采用的多粒度注意力权重超参为 $[0.5,0.3,0.2]$。其他设置与 DETR 保持一致。

4.1.2　评价指标

目标检测任务中常用评价指标为平均正确率（mAP）以及召回率（Recall）。mAP 为各个类别的平均正确率，其各自的正确率计算由在给定 IoU 阈值下计算出来的真正例 TP（True Positive，正样本并且预测正确）和假正例 FP（False Positive，负样本并且预测错误）以及假反例 FN（False Negative，正样本但是预测错误）所决定；Recall 由真正例以及假反例所决定。

DETR 中由于引入了匈牙利算法来对解码器输出进行二分匹配，因此其对目标检测效果的影响也是不能忽略的。其主要损失计算由分类损失、边界框损失组成，而边界框的损失计算又可分为 L1 范数的距离回归损失以及 GIoU 的边框面积损失计算，二者采用不同的权重 λ_{giou} 和 λ_{L1} 对两个损失进行

比例调整，实验中采用 $\lambda_{giou}=5$，$\lambda_{L1}=2$，$\lambda_{cls}=2$。式(28)和式(29)中的 b_m 和 c_m 分别代表 Ground Truth 的边框信息以及类别信息，$\hat{b}_m$ 和 $\hat{c}_m$ 代表模型预测的输出。

$$AP=\frac{TP}{TP+FP} \tag{26}$$

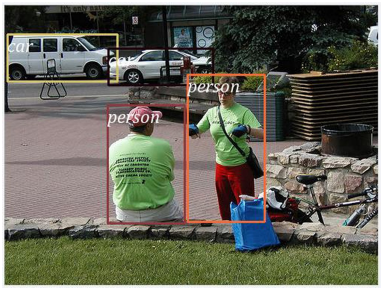
$$Recall=\frac{TP}{TP+FN} \tag{27}$$

$$box=\sum_{m=1}^{M}\lambda_{giou}GIoU(b_m,\hat{b}_{m(i)})+\lambda_{L1}L1(b_m,\hat{b}_{m(i)}) \tag{28}$$

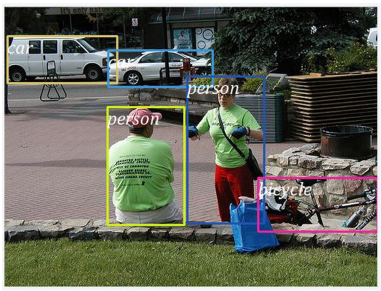
$$loss=\sum_{m=1}^{M}-\lambda_{cls}\log\hat{p}_{\hat{\sigma}(m)}(c_m)+box \tag{29}$$

4.2 改进模型的实验效果对比

图 10(a)展示了未改进的模型对图片进行推理的实验结果，图中共有 5 个待检测目标，显然未改进的模型对检测目标有漏检，例如上边有辆车明显地被漏检了，右下角有辆大部分被遮挡的自行车未被有效检测到，并且模型收敛的回归边界不太准确，偏离了目标原始的位置。而图 10(b)展示了改进的模型检测的结果。可以发现，在经过本文方法改进后，模型的检测性能得到了一定的提升，具体表现在：漏检率降低，所有目标都能被准确识别到，尤其是被遮挡的自行车以及上面的白色的车都被检测到了。



(a)未改进的模型推理结果



(b)改进后的模型推理结果

图 10 改进前后的模型推理结果对比

Fig. 10 Model inference result comparison before and after improvement

4.3 消融实验

为分别验证从解码器输出向量学习到的空间先验(Spatial Prior)对网络模型收敛的影响以及采用多粒度(Multi Granularity)空间注意力机制所带来的准确度提升(DETR-MG)，我们进行了消融实验。图 11 与图 12 分别展示了在加入空间先验和使用多粒度空间注意力与原始 DETR 的训练损失变化曲线和平均准确率的变化曲线，表 1 列出了 3 种网络需要收敛的轮数以及在收敛时的平均准确率(mAP)。

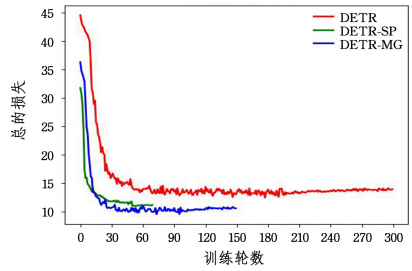


图 11 训练损失变化曲线

Fig. 11 Train loss curve

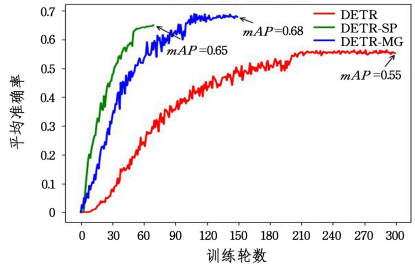


图 12 平均准确率变化曲线

Fig. 12 Average precision curve

表 1 模型收敛所需轮数及其 mAP

Table 1 Epoch number rquired for model convergence and its mAP

模型	收敛轮数	收敛时 $mAP/\%$
DETR	210	55
DETR-SP	60	65
DETR-MG	120	68

4.3.1 空间先验对网络收敛速度的影响

为验证在加入了对解码器输出向量进行参考点学习后整个网络监督训练受到影响，本实验设置总的训练 epoch 为 70，初始学习率为 0.0001，50 个 epoch 后衰减为 10%。从图 11 可以发现，DETR-SP 对物体的综合表现比 ETR 要好很多，最开始的损失相比 DETR 来说更加稳定，下降速度也更快，在最终收敛时，总损失依旧比 DETR 小，而 DETR 在 90 个 epoch 左右，训练损失已经趋近于拟合状态，已经很难再从中学习到更多的特征信息来帮助模型进行检测。图 12 展示了平均准确率变化曲线 mAP，可以看到，随着训练轮次的增加，本文提出的加入空间先验监督的模型在 PASCAL VOC2012 目标检测数据集上准确率上升梯度较大；在进行 60 轮左右的训练后，mAP 曲线趋于平缓，说明网络已经处于收敛状态，而原始的 DETR 在 210 个 epoch 后才开始收敛，其上升曲线比较缓慢。相较而言，DETR-SP 网络训练的收敛速度快了 3.5 倍，并且其平均准确率相比 DETR 提升了近 10%。由此更加证明了在加入对解码器输出嵌入的空间学习后，解码器的交叉注意力模块使得 object query 更容易学习到编码器的输出嵌入，具体来说就是更容易学到内容嵌入上的上下文表示。

4.3.2 多粒度空间注意力对检测效果的影响

和原始的 DETR 中采取的无差别平均注意力计算机制不同，本实验中采用的 DETR-MG(Multi-granularity)根据当前像素点与需要与之进行注意力计算的像素点之间的距离来

决定注意力权重的选取。有 3 个距离等级，距离越近采用更大的注意力权重，距离越远使用更小的注意力权重。实验训练的 epoch 为 150。图 11 描述了模型的训练损失变化曲线，相较于 DETR，DETR-MG 的总训练损失曲线下降速度更快且稳定在一个区间内。图 12 展示了平均准确率变化曲线，DETR-MG 在经过 120 个 epoch 左右开始逐渐收敛，最终 mAP 为 0.68 左右，相较于 DETR 的 0.55，提升了 13%。

4.3.3　空间先验监督与多粒度注意力的共同影响

从上述实验结果可以发现，通过学习解码器输出嵌入的位置特征表达并且将其加入到解码器的交叉注意力模块中作为监督使用，可以大大缩短网络的收敛过程，并且提升模型性能；通过将编码器的自注意力机制以及解码器的交叉注意力替换为本文提出的多粒度空间注意力模块，可以大幅提升模型的检测性能，并且略微加快网络的收敛速度。为进一步验证本文方法的有效性，并且探究空间先验知识与多粒度空间注意力的协同作用，图 13 描述了将两者结合在一起对模型检测性能的影响。训练总损失变化曲线表明了本文方法有助于模型快速学习，降低损失，相较于单独使用两种方法，本文模型在收敛时可以进一步降低总的损失值，表 2 列出了在各种情况下的收敛损失。图 14 展示了同时使用两种方法对模型的平均准确率的影响，可以发现，与单独使用两种方法比较，最后的平均准确率虽然有所提升，但是收敛速度相比只使用空间先验监督反而有所下降，相较于只使用多粒度注意力，收敛速度相近，但准确率有所提高。出现此现象的原因是多粒度空间注意力在解码器的交叉注意力模块对学习到的空间先验知识进行了弱化，对不同的空间位置采用了不同的注意力权重，而这种权重的选取有可能因为数据集中的目标大小与尺寸分布不均匀，进而导致二者不匹配，从而使得空间先验位置信息被减弱。

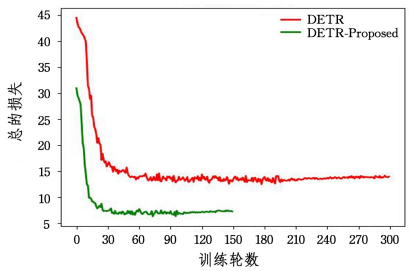


图 13　改进的 DETR 训练损失变化曲线

Fig. 13　Train loss curve of advanced DETR

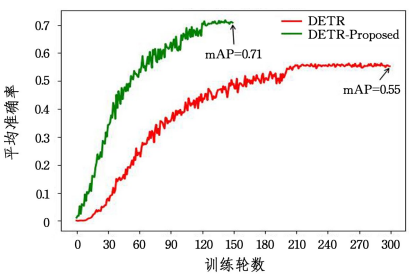


图 14　改进的 DETR 平均准确率变化曲线

Fig. 14　Average precision curve of advanced DETR

表 2　不同配置下的模型性能表现

Table 2　Model performance in different configurations

模型	收敛轮数	收敛时 *mAP*/%	收敛时总损失
DETR	210	55	13.95
DETR-SP	60	65	11.22
DETR-MG	120	68	10.60
DETR-MG-SP	120	71	7.27

4.4　模型对比实验

为充分说明改进方法的有效性，本文选取了 2022 年基于 Transformer 的在目标检测领域使用的 SOTA 模型作为对比。具体包括：采用监督和无监督训练相结合的并能适配于不同分辨率以及多尺度图片的 Swing Transformer V2[20]；通过采样“目标显著点”的特征使 object query 和图像特征的语义对齐，采用这种将特征点和目标语义对齐的匹配方式实现 DETR 快速收敛的 SAM-DETR[21]；通过去噪解决 DETR 预测头做匈牙利算法二分匹配不稳定问题的 DN-DETR[22]。通过将本文方法和上述 3 个基于 Transformer 架构的模型在数据集 PASCAL VOC2012 上的实验结果进行对比，证明了本文改进模型的有效性。表 3 列出了各模型在本实验数据集上的性能表现。

表 3　不同模型的性能表现对比

Table 3　Performance comparison of different models

模型	收敛轮数	收敛时 *mAP*/%
DETR	210	55
DETR-Proposed	60	**71**
Swing Transformer V2	70	68
SAM-DETR	**50**	66
DN-DETR	80	69

从表 3 中可以看出，相较于原来的 DETR 基线，各模型不论是在收敛速度还是收敛时的准确度上都有较大提升。另一方面，虽然 Swing Transformer V2，SAM-DETR，DN-DETR 等 SOTA 模型在 COCO，JFT-300M 等大规模数据集上表现良好，但在较小的公开数据集上的性能表现仍然有待优化，并且本文提出的改进方法相较于其他改进方法性能更好，分别在收敛速度上优于 Swing Transformer V2 和 DN-DETR，在准确率上也优于其他模型。说明在中小规模数据集上本文方法是有效的。而这也说明了另外一个问题，即对于 Transformer 架构而言，大量的数据集以及有监督的训练标签对于 Transformer 架构而言是非常重要的。

4.5　局限性分析

本文方法加入了解码器的空间先验监督模块以及将编解码器的多头注意力模块替换为本文提出的多粒度空间注意力模块，虽然提高了模型整体训练收敛的速度，但也导致了在单次的训练中，计算复杂度有一定的增加，主要表现为：多粒度注意力模块在计算注意力时，需要额外计算当前像素与其他像素的距离并选择对应的权重。另外，由于本方法选取了不同粒度的注意力机制，因此在处理独立的、未经遮挡的目标时有较好效果，但是针对密集目标以及遮挡目标的检测效果不佳，原因主要是本文方法采用的是和原始的 DETR 相同的负样本的采样策略，因此当同一区域有多个目标时，密集目标以及遮挡目标往往容易被错漏或被错判为同一目标，导致效果不佳。

结束语 针对 DETR 中识别精度较低的问题，使用本文提出的多粒度空间注意力替换掉原始的 encoder 中的多头自注意力以及 decoder 中的多头交叉注意力模块，大幅提升了模型的检测能力并在一定程度上加快了收敛速度。针对 DETR 训练收敛慢的问题，提出使用空间先验对网络训练进行监督，限制其学习的空间嵌入信息，进而加快收敛过程。通过上述两个方面对 DETR 模型进行改进，并在 PASCAL VOC2012 目标检测数据集上进行实验分析，结果表明，本文方法相较于原始的 DETR 收敛速度快了 2 倍，识别精度提高了 16%，证明了本文方法的有效性。此外，还做了各个改进部分的消融实验，分别验证了各个改进模块的作用。最后和 2022 年基于 Transformer 架构的 SOTA 目标检测网络在本实验数据集上进行了数据对比分析。

参 考 文 献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[3] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[4] REN S, HE K, GIRSHICK R, et al. Faster r-CNN: Towards real-time object detection with region proposal networks[J]. arXiv: 1506.01497, 2015.

[5] DUAN K, BAI S, XIE L, et al. Centernet: Keypoint triplets for object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6569-6578.

[6] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 734-750.

[7] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition. 2017: 2117-2125.

[8] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768.

[9] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 3-19.

[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations. 2021: 1-22.

[11] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image Transformers & distillation through attention[C]// International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[12] WANG W, XIE E, LI X, et al. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578.

[13] WU H, XIAO B, CODELLA N, et al. Cvt: Introducing convolutions to vision Transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 22-31.

[14] LI Y, CHEN Y P, WANG T, et al. Tokens-to-token vit: Training vision Transformers from scratch on imagenet[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 558-567.

[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[C]// Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, Part I 16. Springer International Publishing, 2020: 213-229.

[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[17] ZHU X, SU W, LU L, et al. Deformable detr: Deformable Transformers for end-to-end object detection[C]// International Conference on Learning Representations. 2021: 1-16.

[18] GAO P, ZHENG M, WANG X, et al. Fast convergence of detr with spatially modulated co-attention[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3621-3630.

[19] YANG J, LI C, ZHANG P, et al. Focal Attention for Long-Range Interactions in Vision Transformers[C]// Advances in Neural Information Processing Systems. 2021: 30008-30022.

[20] LIU Z, HU H, LIN Y, et al. Swin Transformer v2: Scaling up capacity and resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12009-12019.

[21] ZHANG G, LUO Z, YU Y, et al. Accelerating DETR convergence via semantic-aligned matching[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 949-958.

[22] LI F, ZHANG H, LIU S, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13619-1362.

LIAO Junshuang, born in 1999, postgraduate. His main research interests include computer vision, object detection, etc.

TAN Qinhong, born in 1968, associate professor. Her main research interests include embedded system design, Internet of Things technology, etc.

（责任编辑：何杨）