ConvFormer:基于Transformer的视觉主干网络

胡 杰[1,2,3,4],昌敏杰[1,2,3,4*],徐博远[1,2,3,4],徐文才[1,2,3,4]
(1. 武汉理工大学汽车工程学院,湖北武汉 430070;2. 武汉理工大学现代汽车零部件技术湖北省重点实验室,湖北武汉 430070;
3. 武汉理工大学汽车零部件技术湖北省协同创新中心,湖北武汉 430070;
4. 武汉理工大学湖北省新能源与智能网联车工程技术研究中心,湖北武汉 430070)

摘 要: 针对主流Transformer网络仅对输入像素块做自注意力计算而忽略了不同像素块间的信息交互,以及输入尺度单一导致局部特征细节模糊的问题,本文提出一种基于Transformer并用于处理视觉任务的主干网络ConvFormer. ConvFormer通过所设计的多尺度混洗自注意力模块(Channel-Shuffle and Multi-Scale attention,CSMS)和动态相对位置编码模块(Dynamic Relative Position Coding,DRPC)来聚合多尺度像素块间的语义信息,并在前馈网络中引入深度卷积提高网络的局部建模能力. 在公开数据集ImageNet-1K,COCO 2017和ADE20K上分别进行图像分类、目标检测和语义分割实验,ConvFormer-Tiny与不同视觉任务中同量级最优网络RetNetY-4G,Swin-Tiny和ResNet50对比,精度分别提高0.3%,1.4%和0.5%.

关键词: 机器视觉;自注意力;主干网络;Transformer
基金项目: 湖北省重大科技专项(No.2020AAA001,No.2022AAA001)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2024)01-0046-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220735

ConvFormer: Vision Backbone Network Based on Transformer

HU Jie[1,2,3,4], CHANG Min-jie[1,2,3,4*], XU Bo-yuan[1,2,3,4], XU Wen-cai[1,2,3,4]
(1. *School of Automotive Engineering*, *Wuhan University of Technology*, *Wuhan*, *Hubei* 430070, *China*;
2. *Hubei Key Laboratory of Advanced Technology for Automotive Components*, *Wuhan University of Technology*, *Wuhan*, *Hubei* 430070, *China*;
3. *Hubei Collaborative Innovation Center for Automotive Components Technology*, *Wuhan University of Technology*, *Wuhan*, *Hubei* 430070, *China*;
4. *Hubei Research Center for New Energy & Intelligent Connected Vehicle*, *Wuhan University of Technology*, *Wuhan*, *Hubei* 430070, *China*)

Abstract: To solve the problem that the mainstream network based on Transformer only does self-attention computation on the input pixel blocks and ignores the information interaction between different pixel blocks, as well as the blurring of local feature details due to a single input scale, a backbone network based on Transformer and used for processing vision tasks is proposed called ConvFormer. ConvFormer aggregates the semantic information between multi-scale pixel blocks through the designed channel-shuffle and multi-scale attention (CSMS) and dynamic relative position coding (DRPC) modules, as well as introduces deep convolution in the feedforward network to improve the local modeling capability of the network. In the image classification, target detection, and semantic segmentation experiments on public datasets ImageNet-1K, COCO 2017, and ADE20K, ConvFormer-Tiny compares with the optimal networks of the same magnitude RetNetY-4G, Swin-Tiny, and ResNet50 in different vision tasks, the accuracy is improved by 0.3%, 1.4%, and 0.5%.

Key words: machine vision; self-attention; backbone network; Transformer
Foundation Item(s): Major Science and Technology Projects in Hubei Province (No.2020AAA001, No.2022AAA001)

收稿日期:2022-06-27;修回日期:2023-04-11;责任编辑:朱梅玉
*通讯作者:昌敏杰

1 引言

在构建神经网络解决计算机视觉任务时,主干网络(backbone network)用于前端提取图片信息,生成抽象的高层特征用于后续算法处理. 一方面,卷积神经网络(Convolutional Neural Network,CNN)通过卷积层和池化层等具备平移不变性的算子处理图像数据[1],具有局部感受野和权值共享能力,能在硬件上并行学习且避免显式的特征抽取而隐式地从训练数据中学习. 自AlexNet[2]使用卷积核搭建主干网络提取抽象特征并在2012年ImageNet图像分类竞赛中取得第一名之后,卷积神经网络架构通过设计变得更深、更密集且卷积形式更复杂[3~5],并拓展至目标检测、语义分割和目标跟踪[1,6,7]等下游任务. 另一方面,Transformer模型采用编码器-解码器架构和自注意力机制,解决了自然语言处理中循环神经网络无法并行处理的问题[8]. 该架构先将输入的二维词符向量通过线性映射提升向量的表征能力,再将长向量通过3个权值变换矩阵得到查询矩阵query、键矩阵key和值矩阵value,对生成的3个等维度矩阵做点积注意力,得到输出矩阵;多头注意力将生成的矩阵排列为多组,分别输入自注意力模块进行运算,让模型在训练中提取到数据多样性特征且能够并行运算. 自注意力模块的输入具有等效性且没有卷积操作,需要嵌入序列编码来对输入序列进行编号,使模型能够利用输入序列的顺序信息. 编码可分为绝对位置编码和相对位置编码. 绝对位置编码由于简单地使用不同频率正余弦函数[8]生成编码整合在输入序列中得到广泛应用,而相对位置编码为不同输入序列之间的距离. 文献[9]提出二维正弦相对位置编码用于图像分类,并展现相对位置编码嵌入的优势. Transformer模型在经过自注意力模块计算之后,接入前馈神经网络(Feed Forward Network,FFN)对自注意力模块的输出做更高维度的映射变换,最后传入非线性函数来加强模型的表征能力.

Transformer模型中自注意力模块拥有处理全局信息的能力,相比卷积神经网络的有效感受野具备明显优势[10]. 因此,有研究尝试依赖自注意力模块搭建主干网络[11],用于目标检测[12]、图像生成[13]及图像分割[14]等领域并成为了一种趋势. ViT[15](Vision Transformer)将输入图片划分为不重叠的像素块(patch)并展平为二维序列来满足输入形式要求,不引入卷积神经网络直接将Transformer模型用于像素块进行图像分类,但是ViT的输入序列长度与输入图片尺寸成平方关系,在做矩阵计算时会导致计算量和内存消耗剧增. 许多工作在ViT上改进来使其更适合视觉任务[16~22],DeiT[18](Data-efficient image Transformers)引入有效的数据增强方法并提出基于token的知识蒸馏策略引导网络更好地学习;DETR[12](Detection Transformer)提出端到端的新范式,将目标检测问题转为无序集合预测问题;CeiT[22](Convolution-enhanced image Transforme)将卷积模块添加到Transformer前,实现对底层局部信息的提取. 以上改进方法均在视觉处理任务中表现出色,但无法有效关联多尺度像素块间上下文信息. 针对上述问题,本文基于Transformer设计主干网络并用于处理计算机视觉任务,主要研究工作如下.

(1)针对输入像素块尺度单一且各个像素块之间缺乏信息交互的问题,设计高效的多尺度混洗自注意力模块(Channel-Shuffle and Multi-Scale attention,CSMS). CSMS模块将图片划分为多尺度的像素块后,利用分组卷积对不同像素块的键值对做通道混洗,使不同像素块间的查询矩阵对应不同尺度的键矩阵和值矩阵,提升模型的多尺度和长距离建模能力.

(2)提出新的动态相对位置编码模块(Dynamic Relative Position Coding,DRPC),不仅规避了如ViT使用的绝对位置编码会限制输入图片尺寸的劣势,而且能够整合不同像素块的邻域信息来增强模型性能. 然后,在前馈网络中引入深度卷积来弥补自注意力模块所缺失的局部信息.

(3)使用所设计的自注意力模块、位置编码模块和前馈网络搭建通用的视觉主干网络ConvFormer. 在公开数据集ImageNet-1K,COCO 2017和ADE20K上使用本文网络做主要的视觉任务图像分类、目标检测和语义分割实验,与其他主流的同量级Transformer网络进行对比,验证了本文所提视觉主干网络的高效性,并设计了充分的消融实验来验证各个模块的有效性.

2 相关工作

2.1 基于CNN的主干网络

目前,计算机视觉网络的标准范式依然是卷积神经网络. 自2012年AlexNet横空出世并得到当年ILSVRC(Imagenet Large Scale Visual Recognition Challenge)竞赛冠军之后,CNN才成为主流. 更深入、更有效的卷积神经架构陆续出现. VGG[3]通过叠加卷积核加深网络扩大感受野并减少计算量;GoogLeNet[23](GoogLe inception Net)既能保持网络结构的稀疏性,又能利用密集矩阵的高计算性能;ResNet[4](Residual neural Network)设计残差结构解决深层网络梯度消失问题;DenseNet[5](Dense convolutional Network)提升信息和梯度在模型中的传输效率,通过特征重用提升模型性能;EfficientNet[24]证明可以利用复合系数统一缩放模型所有维度来提高模型性能. 除了上述架构上的进步之外,也有许多工作改进单个卷积层,例如深度卷积[25]和可变形卷积[26]. 随着CNN的发展,研究人员还寻求改进CNN的上下文信息,例如ASPP[27](Atrous Spa-

tial Pyramid Pooling)和PPM[28](Pyramid Pooling Module)增强具有多尺度上下文信息的CNN,而NLNet[29](Non-Local neural Networks)和CCNet[30](Criss-Cross Network)为CNN提供了一种非局部机制.

2.2 基于Transformer的主干网络

ViT作为开创性工作通过将图像标记化和展平为标记序列,直接将Transformer架构用于同一大小图像块进行图像分类.随后,基于Transformer架构设计更合理的主干网络来用于处理计算机视觉任务,Swin[16]在固定窗口大小的像素块做自注意力计算降低计算复杂度,同时提出滑动窗口策略弥补目标丢失的部分粗粒度特征;PVT(Pyramid Vision Transformer)[17]引入金字塔结构逐层对特征图进行下采样并使用了空间缩减注意力模块来权衡模型效率和准确率;CAT[19](Cross Attention Transformer)在像素块内交替自注意模块以捕获局部信息,通过在特征图通道方向上划分的不同像素块之间应用注意力捕获全局信息;DPT[20](Deformable Patch-based Transformer)通过参数学习能够自适应变化像素块形状,避免规则划分导致目标完整语义信息分裂至不同像素块中;TNT[21](Transformer in Transformer)细化图片结构,递归使用Transformer,同时关注序列和像素层级的信息提取,显著提高模型对局部结构的建模能力;CrossViT[31]使用两种尺度分别对图像进行划分并独立编码,对编码后的多尺度特征利用交互注意层实现两种尺度序列之间的信息交互.

3 研究方法

3.1 模型架构

本文旨在基于Transformer架构设计自注意力模块、位置编码模块和前馈网络来搭建高效通用的视觉主干网络.如图1(*a*)所示,以ConvFormer分类网络模型进行阐述,输入图像经过4阶段递进提取特征信息后,根据目标类别数接入检测头实现分类功能.图1(*b*)为CSMS模块内部结构,信息流动如下式:

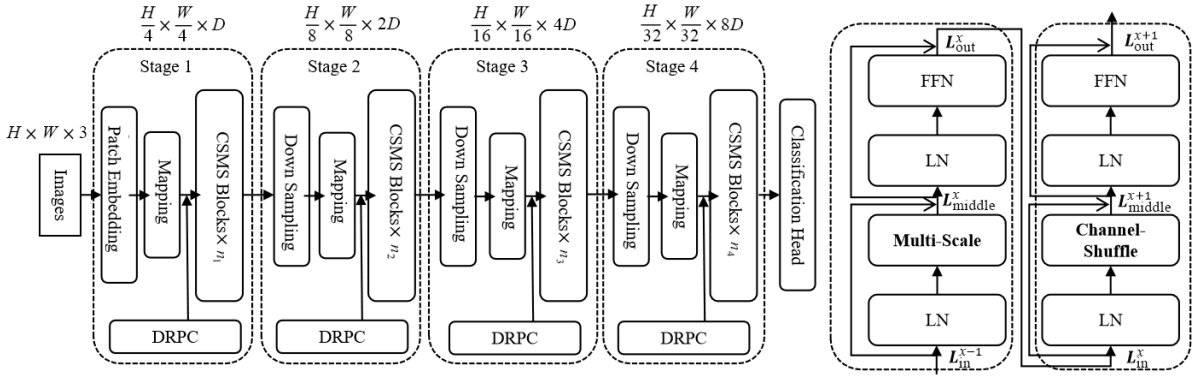
$$\boldsymbol{L}^{x}_{\text{mid}}=\boldsymbol{L}^{x-1}_{\text{in}}+\text{Multi-Scale}(\text{LN}(\boldsymbol{L}^{x-1}_{\text{in}})) \tag{1}$$

$$\boldsymbol{L}^{x}_{\text{out}}=\boldsymbol{L}^{x}_{\text{mid}}+\text{FFN}\left(\text{LN}\left(\boldsymbol{L}^{x}_{\text{mid}}\right)\right)=\boldsymbol{L}^{x}_{\text{in}} \tag{2}$$

$$\boldsymbol{L}^{x+1}_{\text{mid}}=\boldsymbol{L}^{x}_{\text{in}}+\text{Channel-Shuffle}(\text{LN}(\boldsymbol{L}^{x}_{\text{in}})) \tag{3}$$

$$\boldsymbol{L}^{x+1}_{\text{out}}=\boldsymbol{L}^{x+1}_{\text{mid}}+\text{FFN}(\text{LN}(\boldsymbol{L}^{x+1}_{\text{mid}})) \tag{4}$$

其中,$\boldsymbol{L}^{x-1}_{\text{in}}\in M^{N\times d}$为经过像素块嵌入模块序列化的二维矩阵,$M$表示多维矩阵序列,LN(·)为层归一化LayerNorm[32],FFN(·)为前馈神经网络,Multi-Scale(·)和Channel-Shuffle(·)组合为本文设计的多尺度混洗自注意力CSMS模块,详细处理步骤如下.



(*a*) ConvFormer分类网络整体架构 (*b*) CSMS块内部结构

图1 模型框架图

为满足自注意力模块输入需求,阶段1像素块嵌入(patch embedding)模块将输入图片规则划分为不重叠等大小的三维像素块记作$\boldsymbol{\delta}\in M^{H\times W\times C}$,通过映射变换(mapping)到更高维度并展平为二维矩阵记作$\boldsymbol{\theta}\in M^{N_\theta\times d_\theta}$.后3个阶段利用下采样(down sampling)模块对输入像素块进行2×2卷积下采样,减少像素块数目并升维形成金字塔结构,各阶段输出维度信息列于图1(*a*)上方.

随后,对每个阶段的输入序列嵌入动态相对位置编码DRPC模块提供像素块位置信息.叠加多尺度混洗自注意力CSMS模块,将输入序列扩充为多尺度像素块提高表征能力,并通过分组卷积做通道混洗操作,打乱键值对本身的语义信息使不同像素块产生上下文交互,提高网络长距离建模能力.单个阶段信息传输过程如下:

$$M_{\text{Stage}}=\begin{cases}\boldsymbol{\delta}=\text{PatchEmbedding}\left(\text{Input}\right)\\ \boldsymbol{\theta}=\text{Mapping}\left(\boldsymbol{\delta}\right)\\ \text{Output}=\text{CSMS}_m\left(\boldsymbol{\theta}+\text{DRPC}\right)\end{cases} \tag{5}$$

其中,Mapping(·)=relu(Linear(·))表示非线性的高维序列映射.

在网络尾部根据数据集类别数插入全连接层作为检测头进行图像分类,其中n=numble of classes为类别数.

$$\text{Detector}=\text{Linear}_n(M_{\text{Stage1,2,3,4}}) \tag{6}$$

依据1:2:4的参数量设计规则搭建3种不同尺度大小的模型即ConvFormer-Tiny,ConvFormer-Medium和ConvFormer-Base. 各模型及网络结构如表1所示.

表1　网络结构参数表

输出尺寸	模块名称	ConvFormer-Tiny	ConvFormer-Medium	ConvFormer-Base
Stage-1 56×56	像素块嵌入模块	卷积核大小:4×4　卷积核步长:4		
	动态相对位置编码模块	卷积核大小:1×1,3×3		
	多尺度混洗自注意力模块	$\begin{bmatrix}D=64,H=2,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 1$	$\begin{bmatrix}D=64,H=2,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 2$	$\begin{bmatrix}D=64,H=2,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 3$
Stage-2 28×28	下采样模块	卷积核大小:2×2　卷积核步长:2		
	动态相对位置编码模块	卷积核大小:1×1,3×3		
	多尺度混洗自注意力模块	$\begin{bmatrix}D=128,H=4,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 2$	$\begin{bmatrix}D=128,H=4,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 4$	$\begin{bmatrix}D=128,H=4,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 4$
Stage-3 14×14	下采样模块	卷积核大小:2×2　卷积核步长:2		
	动态相对位置编码模块	卷积核大小:1×1,3×3		
	多尺度混洗自注意力模块	$\begin{bmatrix}D=256,H=8,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 4$	$\begin{bmatrix}D=256,H=8,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 10$	$\begin{bmatrix}D=256,H=8,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 20$
Stage-4 7×7	下采样模块	卷积核大小:2×2　卷积核步长:2		
	动态相对位置编码模块	卷积核大小:1×1,3×3		
	多尺度混洗自注意力模块	$\begin{bmatrix}D=512,H=16,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 1$	$\begin{bmatrix}D=512,H=16,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 1$	$\begin{bmatrix}D=512,H=16,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 2$

表1中,$\begin{bmatrix}D=64,H=2,\\ \boldsymbol{K}_M=[2,4,8],G_s=3\end{bmatrix}\times 1$表示该Stage有1个CSMS模块,输入通道数为64,多头注意力个数为2,多尺度卷积操作中卷积核大小为2×2,4×4和8×8,通道混洗中分组数为3.

3.2　多尺度混洗自注意力CSMS模块

针对主流Transformer网络输入像素块尺度单一且无法在不同像素块之间产生信息交互的问题,本文在主干网络ConvFormer中设计多尺度混洗自注意力CSMS模块来替换ViT中的传统自注意力模块. CSMS模块包括多尺度信息提取Multi-Scale模块和通道混洗Channel-Shuffle模块. Multi-Scale模块旨在生成多尺度序列捕获多尺度目标的语义信息. Channel-Shuffle模块旨在打乱查询矩阵的键值对. CSMS模块对不同像素块上下文信息建模,模块整体结构如图2所示.

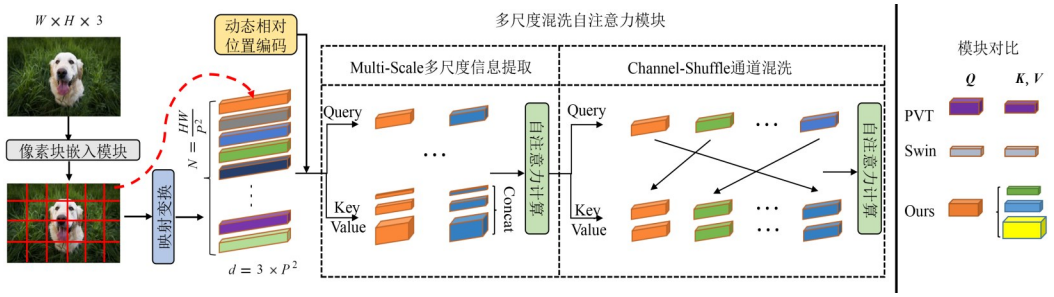


图2　CSMS模块整体图

首先,将输入图像$\boldsymbol{x}\in M^{H\times W\times 3}$划分为不重叠等大小为$P$的像素块,序列化为自注意力模块所需要的二维矩阵形式$\boldsymbol{L}_x\in M^{N_x\times d_x}$并嵌入位置编码DRPC模块,其中,$N_x=\dfrac{H\times W}{P^2}$,$d_x=3\times P^2$,输出记作$\boldsymbol{\gamma}$,即

$$\boldsymbol{\gamma}=\boldsymbol{L}_x+\text{DRPC} \tag{7}$$

$$\boldsymbol{L}_x=\text{Mapping}(\text{PatchEmbedding}(\boldsymbol{x})) \tag{8}$$

然后,将输出$\boldsymbol{\gamma}$传入多尺度信息提取Multi-Scale模块. 在卷积神经网络中,Faster RCNN[1]设计的多尺度锚点只依赖单一尺度的图像、滤波器和特征映射来预测多尺度目标,FPN[33](Feature Pyramid Networks)特征金字塔网络对自上而下的每一层都上采样来获取更准确的像素位置信息,通过特征融合来更加精准地检测多尺度目标;Transformer前沿网络PVT[16]和Swin[17]将输入图片划分为等大小的像素块并序列化为二维矩阵时,忽视输入特征图尺度对像素块尺度的影响,使划分

的像素块尺度单一而丢失目标的多尺度语义信息，导致模型性能降低. 因此，Multi-Scale模块在底层针对输入构建多尺度像素块而不是简单地使用多尺度卷积核来处理每一个像素点，在Transformer编码阶段充分利用输入信息，使网络获得多尺度特征提升性能.

Multi-Scale模块具体如图3所示. 将特征图 $\boldsymbol{I}\in M^{H\times W\times C}$ 进行等大小像素块划分后得到伪特征图 $\boldsymbol{I}_P\in M^{H'_P\times W'_P\times C}$，其中，$H'_P=\dfrac{H}{H_P}$，$W'_P=\dfrac{W}{W_P}$. 序列化为二维矩阵前使用2×2和4×4的卷积核对伪特征图进行多尺度特征提取并序列化生成两对键值对 $[\boldsymbol{K}_1,\boldsymbol{V}_1]\in M^{N_1\times C_\alpha}$ 和 $[\boldsymbol{K}_2,\boldsymbol{V}_2]\in M^{N_2\times C_\alpha}$，查询矩阵在通道方向拆分为 $[\boldsymbol{Q}_1,\boldsymbol{Q}_2]\in M^{N\times C_\alpha}$，其中，$N=H_P\times W_P$，$N_1=\dfrac{H_P\times W_P}{4}$，$N_2=\dfrac{H_P\times W_P}{16}$，$C_\alpha=\dfrac{C}{2}$. 对 $[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}]$ 做自注意力矩阵运算，自注意力计算公式如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{9}$$

分别得到包含多尺度信息的二维序列 $[\boldsymbol{L}_1,\boldsymbol{L}_2]\in M^{N_L\times C_L}$，其中，$N_L=H_P\times W_P$，$C_L=\dfrac{C}{2}$，将序列三维化后在通道方向拼接得到相同维度且包含多尺度语义信息的像素块输出 $\boldsymbol{I}^1_{\text{out}}\in M^{H_P\times W_P\times C}$，即

$$\boldsymbol{I}^1_{\text{out}}=\text{Concat}(\boldsymbol{L}_1+\boldsymbol{L}_2) \tag{10}$$

$$\boldsymbol{L}_{1,2}=\text{Attention}\left([\boldsymbol{Q}_1,\boldsymbol{Q}_2],[\boldsymbol{K}_1,\boldsymbol{K}_2],[\boldsymbol{V}_1,\boldsymbol{V}_2]\right) \tag{11}$$

$$\boldsymbol{Q}=\text{Split}(\boldsymbol{I}) \tag{12}$$

$$[\boldsymbol{K},\boldsymbol{V}]=\text{Conv}(\text{PatchEmbedding}(\boldsymbol{I})) \tag{13}$$

最后，将Multi-Scale模块的输出 $\boldsymbol{I}^1_{\text{out}}$ 传入Channel-Shuffle模块. 点乘注意力无法关注像素块间相关信息[16]，Swin设计卷积滑动窗口对像素块进行滑动建模，但依旧无法连接远距离序列且模板性太强不利于网络学习. ShuffleNet[34]采用通道打乱能避免常规分组卷积产生的边界效应. 基于此思想，Channel-Shuffle模块引入学习参数对输入序列分组后进行通道混洗操作，在不同像素块间键值对产生信息流动，提高网络的长距离建模和特征学习能力.

如图3所示，将 $\boldsymbol{I}^1_{\text{out}}$ 序列化为二维矩阵后生成对应的查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$，然后将键值对转换为三维特征图，Shuffle(·)表示在通道方向分组后提取对应位置的通道并合并成新的键值对，使每个像素块都能共享其他像素块语义特征，再与查询矩阵做自注意力计算，最终输出 $\boldsymbol{I}^2_{\text{out}}\in M^{H_P\times W_P\times C}$ 的像素块之间产生信息交互且具有多尺度语义信息，即

$$\boldsymbol{I}^2_{\text{out}}=\text{reshape}(\text{Attention}(\boldsymbol{Q},\boldsymbol{K}',\boldsymbol{V}')) \tag{14}$$

$$[\boldsymbol{K}',\boldsymbol{V}']=\text{shuffle}(\boldsymbol{K},\boldsymbol{V}) \tag{15}$$

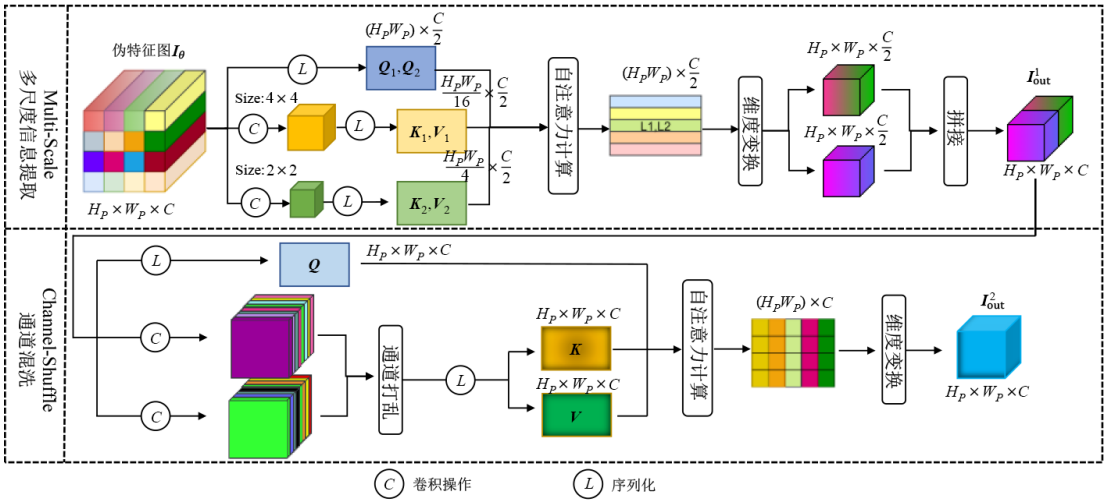


图3 CSMS模块信息流动图

如图2右侧的对比部分所示，与PVT[16]和Swin[17]中Query对应单一尺度的Key和Value不同，CSMS模块输出的查询矩阵能匹配乱序像素块中的多尺度键值对. 因此，CSMS模块不仅可以聚合多个像素块间的语义信息，而且能够捕获目标的多尺度特征信息. 以上为CSMS模块完整计算流程. 与其他Transformer网络同时处理Query，Key和Value不同，ConvFormer仅关注键值对来降低部分计算复杂度且模型性能表现更佳.

3.3 动态相对位置编码DRPC模块

针对传统位置编码会限制输入尺寸及像素块的划分方式的问题，本文提出DRPC模块，自适应变化维度且引入像素块的邻域信息保证其有序性. 如图4右所示ViT采用绝对位置编码方式，将二维序列像素块 $\boldsymbol{P}\in M^{N\times d}$ 和位置编码 $\boldsymbol{P}_{\text{coding}}\in M^{N\times 1}$ 在序列方向上拼接记作 $\boldsymbol{Y}_{\text{ViT}}\in M^{N\times d_c}$，其中，$d_c=d+1$，即

$$\boldsymbol{Y}_{\text{ViT}}=\text{Concat}(\boldsymbol{P}+\boldsymbol{P}_{\text{coding}}) \tag{16}$$

因此,ViT受固定大小像素块约束,嵌入位置编码长度必须与像素块长度对应,导致只能处理预设大小的图片.本文提出的动态相对位置编码DRPC模块不受输入图片和划分像素块尺寸的限制,使用卷积生成抽象特征图作为编码模块嵌入像素块,能随着输入像素块数量变化而动态改变其维度,嵌入Transformer架构中训练产生的参数量可忽略不计,具体细节如图4所示.

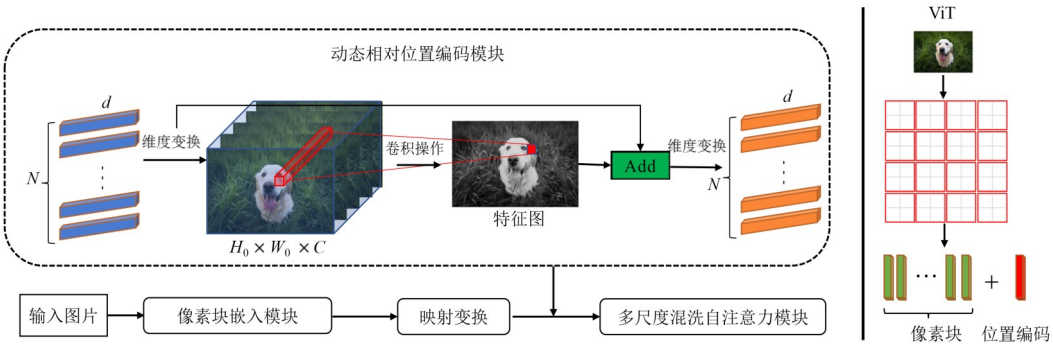


图4　DRPC模块内部结构

将维度为$N \times d$的二维矩阵逆变换为$C \times H \times W$的图像,并通过卷积核生成同样维度$C \times H \times W$的过渡特征图Feature Map作为编码机制DRPC模块,取代传统ViT通过正余弦函数随机生成的绝对位置编码.此时,DRPC模块动态随输入尺寸变化其大小,始终与输入图像维度一致,从而摆脱ViT只能对固定长度序列编码的问题,而且对不同像素块的邻域信息进行建模,使权值更大的像素块之间联系更紧密.

利用残差结构将变换后等维度的特征图与原始图像相加记作$\boldsymbol{Y}_{\text{Conv}} = M^{H \times W \times C}$,聚合DRPC模块中的位置信息和初始目标的语义信息,即

$$\boldsymbol{Y}_{\text{Conv}} = \text{Add}\left(\text{Res}(\boldsymbol{x}) + \text{DRPC}\right) \tag{17}$$

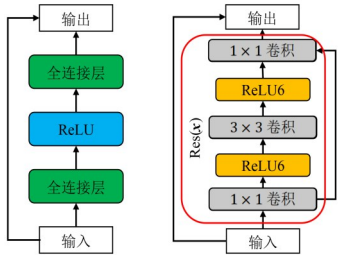
最后,将特征图逆变换为二维矩阵形式传入自注意力CSMS模块进行计算.与广泛使用的绝对和相对位置编码不同,即插即用的DRPC模块能自适应输入尺寸变化及不同大小像素块的划分方式,在各个层次引入DRPC模块扩大像素块的感受野,并使其对应的编码点具备领域信息,保持图像分类任务中所需的平移不变性,并提供一定程度目标的绝对位置编码信息用于目标检测任务.

为了在自注意力模块中整合卷积神经网络处理局部信息的能力来弥补细粒度特征,如图5所示,本文与ViT中的前馈网络简单采用全连接层不同,参考MobileNet[35]的设计架构引入1×1和3×3卷积在不增加参数量的前提下加深网络,并增强网络提取特征图局部信息的能力.将激活函数ReLU替换为ReLU6函数,减少参数大小并让模型更早地学到稀疏特征,并添加残差分支使前馈网络便于训练[4].

FFN结构信息流动如下:

$$\text{FFN}_{\text{old}} = \boldsymbol{x} + \text{Linear}(\text{relu}(\text{Linear}(\boldsymbol{x}))) \tag{18}$$

$$\text{FFN}_{\text{new}} = \boldsymbol{x} + \text{Res}(\boldsymbol{x}) \tag{19}$$



(a) ViT　　(b) Ours

图5　前馈网络FFN对比

4　实验

本文设计的视觉主干网络ConvFormer在公开数据集ImageNet-1K[36],COCO 2017[37]和ADE20K[38]上做图像分类、目标检测和语义分割对比试验.与前沿主干网络进行比较,包括卷积神经网络(ResNet[4],ResNeXt[25],RegNetY[39])以及Transformer网络(Swin[16],PVT[17],DeiT[18],CAT[19],DPT[20],TNT[21],CeiT[22],CrossViT[31]),利用消融实验验证各个模块的有效性,通过充分的实验验证所提网络的有效性,实验环境如表2所示.

表2　实验环境

	环境设置
操作系统	Ubuntu18.04
GPU	NVIDIA Tesla V100 32 GB×4
CPU	Intel Xeon Glod 6248R 3.0 GHz 24核48线程
软件	Python3.6.9, PyTorch1.7.0, CUDA10.2, CuDNN7.6.5

4.1　图像分类

图像分类实验使用的ImageNet-1K数据集包含来自1 000个类别的128万张训练图片和5万张验证图片,在训练集上训练模型,并用验证集测试输出排名第一的类别与实际结果相符的准确率(Top-1 accuracy,

Top-1). 训练时选择初始学习率为0.001并使用了衰减权重为0.05余弦衰减的优化器AdamW[40],将动量设置为0.9. 数据增强方法采用随机裁剪,随机翻转和标签平滑,正则化方法采用Cutmix[41]和Mixup[42]. 所有模型都在4张NVIDIA Tesla V100显卡上以128的批次(batch-size)从头开始训练300个轮次(epoch),所有图片都裁剪为统一的224×224像素进行训练和测试.

图像分类实验结果如表3所示,ConvFormer较其他模型性能更优. 与同量级Top-1精度最高的网络对比,ConvFormer-Tiny、ConvFormer-Medium 和 ConvFormer-Base在较RegNetY-4G,CAT-S和Swin-S参数量分别降低34.3%,34.9%和14.4%的前提下,精度提高0.3%,0.4%和0.3%,验证了ConvFormer作为通用视觉主干网络的潜力.

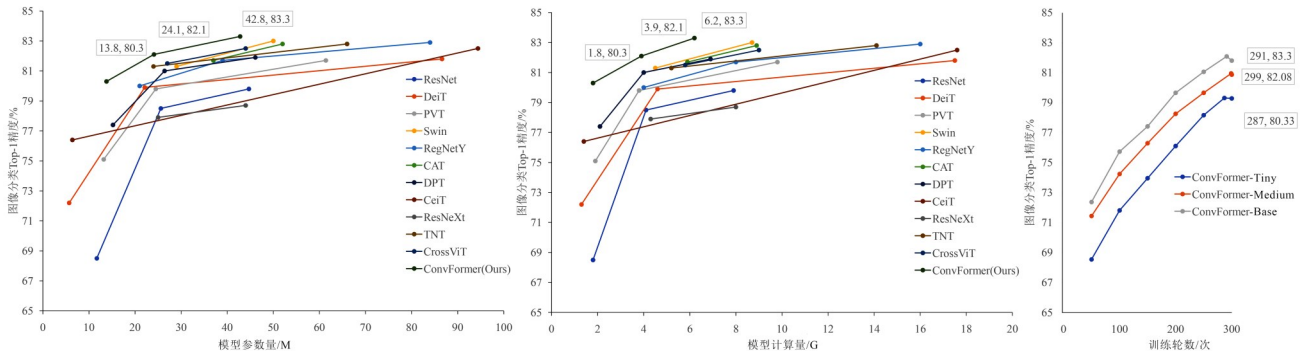
一方面,多尺度混洗自注意力CSMS模块较其他对比网络中设计的自注意力模块,通过多尺度卷积和通道混洗机制,能交互多个尺度像素块间的上下文信息以增强网络长距离建模能力,使得搭建的主干网络有较强的特征提取能力,从而分类精度更高;另一方面,通过对自注意力模块键值对、位置编码模块及前馈网络的简化表征,使得网络参数量低于其他网络.

表3 图像分类实验结果表

主干网络	参数量/M	计算量/G	Top-1/%	参考来源	方法
R18[4]	11.7	1.8	68.5	CVPR 2016	CNN
DeiT-T[18]	5.7	1.3	72.2	ICMR 2021	Transformer
PVT-T[17]	13.2	1.9	75.1	ICCV 2021	Transformer
CeiT-T[22]	6.4	1.4	76.4	ICCV 2021	Transformer
DPT-T[20]	15.2	2.1	77.4	ACM 2021	Transformer
RegNetY-4G[39]	21.0	4.0	80.0	CVPR 2020	CNN
ConvFormer-Tiny	13.8	1.8	**80.3(+0.3)**	—	Transformer
ResNeXt-50[25]	25.0	4.3	77.9	CVPR 2017	CNN
R50[4]	25.6	4.1	78.5	CVPR 2016	CNN
PVT-S[17]	24.5	3.8	79.8	ICCV 2021	Transformer
DeiT-S[18]	22.1	4.6	79.9	ICMR 2021	Transformer
DPT-S[20]	26.4	4.0	81.0	ACM 2021	Transformer
TNT-S[21]	24.0	5.2	81.3	NIPS 2021	Transformer
Swin-T[16]	29.0	4.5	81.3	ICCV 2021	Transformer
CrossViT-15[31]	27.0	5.8	81.5	ICCV 2021	Transformer
RegNetY-8G[39]	39.0	8.0	81.7	CVPR 2020	CNN
CAT-S[19]	37.0	5.9	81.7	ICME 2022	Transformer
ConvFormer-Medium	24.1	3.9	**82.1(+0.4)**	—	Transformer
ResNeXt-101[25]	44.0	8.0	78.7	CVPR 2017	CNN
R101[4]	44.7	7.9	79.8	CVPR 2016	CNN
PVT-M[17]	44.2	6.7	81.2	ICCV 2021	Transformer
PVT-L[17]	61.4	9.8	81.7	ICCV 2021	Transformer
DeiT-B[18]	86.6	17.5	81.8	ICMR 2021	Transformer
DPT-M[20]	46.1	6.9	81.9	ACM 2021	Transformer
CrossViT-18[31]	44.0	9.0	82.5	ICCV 2021	Transformer
CeiT-B[22]	94.4	17.6	82.5	ICCV 2021	Transformer
CAT-B[19]	52.0	8.9	82.8	ICME 2022	Transformer
TNT-B[21]	66.0	14.1	82.8	NIPS 2021	Transformer
RegNetY-16G[39]	84.0	16.0	82.9	CVPR 2020	CNN
Swin-S[16]	50.0	8.7	83.0	ICCV 2021	Transformer
ConvFormer-Base	42.8	6.2	**83.3(+0.3)**	—	Transformer

图6(a)和图6(b)为各模型在不同参数及计算量下的Top-1检测结果折线图,从图中可以更直观地看出同参数量下本文提出的模型ConvFormer全面优于其他对比模型. 图6(c)表明ConvFormer训练趋势平稳且没有出现过拟合的情况.

为观察自注意力模块的作用方式,将模型最后一层全连接层输出各个像素的分类得分映射回原图,得到不同大小模型自注意力模块可视化图. 如图7所示,



(*a*) 模型参数量　　(*b*) 模型计算量　　(*c*) 训练轮数

图6　实验结果折线图

3个不同大小的ConvFormer模型均能较好地捕捉物体的全局信息. ConvFormer-Base因为模型层数更深、自注意力块叠加更多,学习到的特征信息更加丰富,较其他两个模型能更加精准地定位图像中的目标.

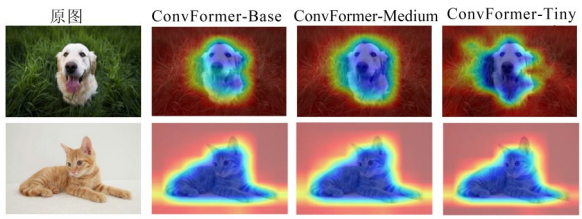


图7　自注意力热力图

4.2　目标检测

目标检测实验使用的COCO 2017数据集包含11.8万张训练图片、5千张验证图片和4.1万张测试图片. 模型在训练集和验证集上进行训练和调参,并在测试集上测试其平均精度(Mean Average Precision, MAP). 本文在标准检测器Mask R-CNN[6]和RetinaNet[43]上评估主干网络. 训练时使用第4.1节中ImageNet图像分类实验得到的训练权重初始化主干网络,并利用Xavier初始化新添加的网络层. 模型Batch-Size设置为2,在4张NVIDIA Tesla V100显卡上训练12个epoch,选择衰减权重为0.05的优化器AdamW,并将初始学习率设为0.000 1,训练和测试的图片尺寸设置为1 300 pixel×800 pixel.

目标检测实验结果如表4和表5所示. 与同量级精度最高网络对比,在COCO目标检测实验中,本文所设计的模型ConvFormer-Tiny在Mask-RCNN评价基准较Swin-T参数量降低27.9%前提下,平均精度MAP提高1.4%,且单帧图片推理耗时63.5 ms,较ResNet50减少9.7 ms;在RetinaNet评价基准较PVT-S参数量降低30.7%前提下,平均精度MAP提高0.4%,且单帧图片推理耗时49.1 ms,较ResNet50减少6.8 ms. 这说明本文所提网络满足实时性需求,在减少参数量和计算量的同时提高检测精度及推理效率.

表4　Mask R-CNN基准下目标检测实验结果表

实验方法	主干网络	参数量/M	mAP^b/%	AP$^b_{50}$/%	AP$^b_{75}$/%	mAP^m/%	AP$^m_{50}$/%	AP$^m_{75}$/%	速度/ms
Mask R-CNN 1×schedule	ResNet50[4]	44	38.0	58.6	41.4	34.4	55.1	36.7	73.2
	PVT-S[17]	44	40.4	62.9	43.8	37.8	60.1	40.3	81.4
	CAT-S[19]	57	41.6	65.1	45.4	38.6	62.2	41.0	87.6
	Swin-T[16]	48	42.2	64.6	46.2	39.1	61.6	42.0	83.1
	ConvFormer-Tiny	**34.6**	**43.6(+1.4)**	64.8	47.9	39.6	61.9	42.4	**63.5(−9.7)**

表5　RetinaNet基准下目标检测实验结果表

实验方法	主干网络	参数量/M	mAP^b/%	AP$^b_{50}$/%	AP$^b_{75}$/%	mAP$_S$/%	mAP$_M$/%	mAP$_L$/%	速度/ms
RetinaNet 1×schedule	ResNet50[4]	37.7	36.3	55.3	38.6	19.3	40.0	48.8	55.9
	ResNet101[4]	57	38.5	57.8	41.2	21.4	42.6	51.1	69.2
	CAT-S[19]	47	40.1	61.0	42.6	24.9	43.6	52.8	82.4
	PVT-S[17]	34.2	40.4	61.3	43.0	25.0	42.9	55.7	76.9
	ConvFormer-Tiny	**23.7**	**40.8(+0.4)**	60.3	43.9	26.0	44.8	53.3	**49.1(−6.8)**

图8为PVT-Small和本文ConvFormer-Tiny在随机挑选复杂场景下的COCO 2017数据集图片且基准同样为RetinaNet下的目标检测对比图. ConvFormer较PVT出现漏检、错检及重复检测的概率要更低,说明CSMS自注意力模块能够有效增强长距离建模能力,且在前馈网络中引入卷积神经网络能保留目标的细粒度特征,同时表

明DRPC模块能对输入像素块高效编码并提供准确的位置信息来进一步提升网络性能. 因此ConvFormer面对复杂环境下的目标检测工作比其他主流模型表现得更好.

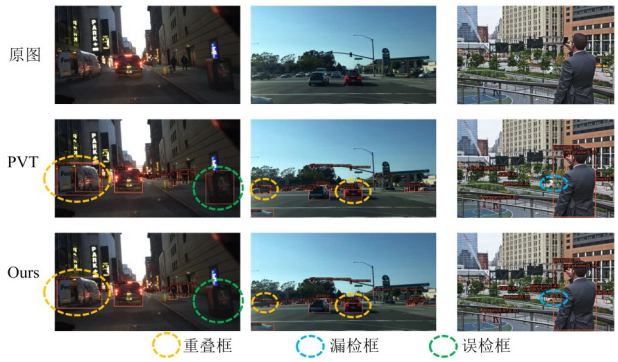


图8 COCO 2017数据集场景目标检测

为了进一步测试模型在复杂场景下的泛化能力,利用如图9所示试验车E70上搭载的工业相机GIGE Camera采集到的校园数据场景进行测试,使用以RetinaNet为基准的ConvFormer-Tiny目标检测模型对输入图片进行推理测试.

检测效果如图10所示. 无论是行人、车辆还是标识牌,ConvFormer均能准确检测到物体轮廓,表明模型具有一定泛化性并能较好地捕捉复杂场景下的多尺度目标.

图9 实验车平台

图10 校园数据场景目标检测

4.3 语义分割

语义分割实验使用的ADE20K数据集包含来自150个语义类别的2万张训练图片、2千张验证图片和3千张测试图片. 模型在训练集和验证集上进行训练和调参,在测试集上测试其真实值和预测值两个集合的交集和并集之比(Mean Intersection over Union,MIoU). 同样使用在ImageNet上预训练得到的权重初始化主干网络并利用Xavier初始化新添加的网络层,使用MMSegmentation[44]中的FPN[33]网络作为基础框架,将Batch-Size设置为4,在4张NVIDIA Tesla V100显卡上迭代80K次. 对于FPN使用初始学习率为0.000 1且权重衰减为0.9的AdamW优化器,输入尺寸设置为512×512. 采用MMSegmentation中默认的数据增强方法,即对图片随机水平翻转、随机缩放和随机光度失真.

语义分割实验结果如表6所示. 以FPN为基准时,ConvFormer-Tiny较CAT-T,PVT-T和ResNet50的MIoU精度指标分别提高1.3%,0.6%和0.5%;单帧图片推理耗时37.5 ms,较CAT-T减少8.2 ms.

表6 语义分割实验结果表

实验方法	主干网络	参数/M	MIoU/%	速度/ms
Semantic FPN	CAT-T[19]	17.3	35.9	45.7
	PVT-T[17]	17.9	36.6	46.1
	ResNet50[4]	29	36.7	45.8
	ConvFormer-Tiny	18.3	37.2(+0.5)	37.5(−8.2)

图11为在ADE20K数据集中随机挑选的图片下,PVT-Small和ConvFormer-Tiny以FPN作为基准的语义分割效果对比. ConvFormer对大目标物体处理的更加平滑,对小目标物体漏检误检概率更小. 因为CSMS模块引入多尺度像素块,所以较其他网络能有效识别不同尺度目标.

为进一步验证ConvFormer分割能力的泛化性,利用校园数据场景进行语义分割测试. 测试效果如图12所示. ConvFormer能准确分割出行人及车辆等不同尺度的物体,说明CSMS自注意力模块能同时捕获目标的粗粒度和细粒度特征. 因此,ConvFormer能准确对复杂环境下多尺度物体进行像素级分割.

4.4 消融实验

为验证主干网络ConvFormer各模块的有效性,本文基于ConvFormer-Tiny设计了如下消融实验. 训练参数设置与图像分类实验保持一致.

(1)将CSMS自注意力模块分别替换为单独等数量的多尺度聚合模块(Channel-Shuffle attention,CS)和通

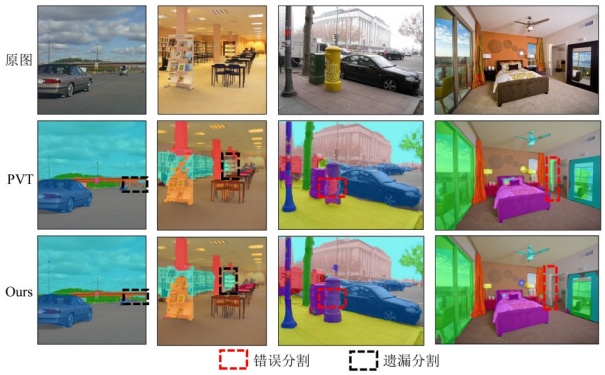


图11　ADE20K数据集场景语义分割

图12　校园数据场景语义分割

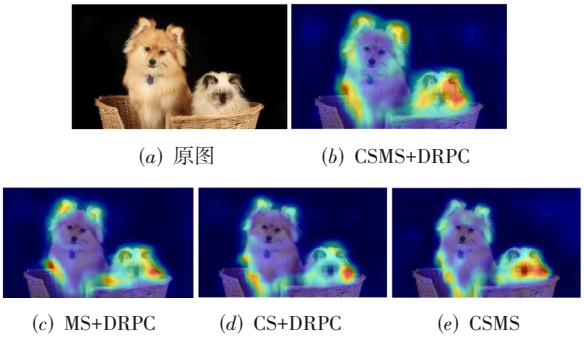
道混洗(Multi-Scale attention,MS)模块.前者取消每个阶段的多尺度输入,仅在划分为多个像素块之后对键值对做通道混洗.如表7所示,实验精度降低1.5%,说明多尺度嵌入模块对整合像素块不同尺度语义信息的能力有较大提升.后者输入为多尺度像素块,但是取消键值对的通道混洗,实验精度降低0.7%,表明采用多通道混洗操作能使不同像素块产生信息交互,从而提高模型长距离建模能力.

(2)将DRPC动态位置编码模块及前馈网络FFN替换为PVT中的绝对位置编码方式及前馈传播方式.如表7所示,结果降低0.6%,验证了本文所设计的编码方式能通过整合像素块的邻域信息有效为像素块提供精确的位置编码信息,从而增强网络的长距离建模能力,以及在前馈网络中聚合卷积神经网络来加深局部特征信息的有效性.

表7　基于ConvFormer-Tiny的图像分类消融实验结果表

多尺度聚合模块	通道混洗模块	动态位置编码模块	Top-1/%
√	√	√	80.3
√	√		79.7
√		√	78.8
	√	√	79.6

利用消融实验设计的3种模型分别加载训练得到的权重文件来处理多尺度图片,利用RetinaNet检测头将网络4个阶段输出的特征图叠加并映射回原图得到多尺度检测目标的热力图.如图13所示,颜色越深表明权重越大关注度越高.图13(*a*)为原始图片;图13(*b*)为使用本文所设计的自注意力和位置编码模块;图13(*d*)由于取消了多尺度嵌入,在处理多尺度目标时会过分关注某个目标而无法平衡粗粒度和细粒度特征;图13(*c*)和图13(*e*)分别为取消通道混洗操作和动态位置编码DRPC模块对应的热力图,两者都会导致不同像素块之间缺乏信息交互而丢失部分类似耳朵和轮廓的上下文信息.



(*a*) 原图　(*b*) CSMS+DRPC

(*c*) MS+DRPC　(*d*) CS+DRPC　(*e*) CSMS

图13　消融实验热力图

5　结束语

本文提出了一种基于Transformer的视觉主干网络ConvFormer,考虑到主流Transformer主干网络忽视了不同像素块的信息交互以及缺少输入像素块多尺度语义特征的问题,设计了CSMS自注意力模块.通过多尺度特征提取以及键值对混洗操作,不仅在不同像素块间建立联系,而且提高了网络模型捕捉多尺度目标语义信息的能力;利用DPRC动态位置编码取代绝对位置编码,通过整合邻域信息增强模型的长距离建模能力并且可以动态适应输入图片大小;通过在前馈网络中整合卷积神经网络局部建模的优势来丰富特征图细粒度信息.使用本文所提出的模型在公开数据集ImageNet-1K,COCO 2017和ADE20K上分别进行图像分类、目标检测和语义分割实验.结果显示,ConvFormer作为主干网络在多项视觉任务中均优于其他对比的同量级网络,本文模型作为主干网络能有效处理各种视觉任务并有望成为计算机视觉任务中的通用主干网络.此外,如何引入偏移参数划分像素块来自适应目标特征以避免割裂导致语义信息被破坏,并降低模型训练时的超参数敏感性,提高模型训练鲁棒性,从而进一步提高模型性能,将是笔者后续的研究方向.

参考文献

[1] REN S Q, HE K M, 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.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2022-06-22]. https://arxiv.org/abs/1409.1556.

[4] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[5] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2261-2269.

[6] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2980-2988.

[7] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]//European Conference on Computer Vision. Cham: Springer, 2016: 850-865.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[9] BELLO I, ZOPH B, LE Q, et al. Attention augmented convolutional networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 3285-3294.

[10] 田永林, 王雨桐, 王建功, 等. 视觉Transformer研究的关键问题: 现状及展望[J]. 自动化学报, 2022, 48(4): 957-979.
TIAN Y L, WANG Y T, WANG J G, et al. Key problems and progress of vision transformers: The state of the art and prospects[J]. Acta Automatica Sinica, 2022, 48(4): 957-979. (in Chinese)

[11] BELTAGY I, PETERS M E, COHAN A. Longformer: The long-document transformer[EB/OL]. (2020-04-10) [2022-06-22]. https://arxiv.org/abs/2004.05150.

[12] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[13] JIANG Y F, CHANG S Y, WANG Z Y. TransGAN: Two pure transformers can make one strong GAN, and that can scale up[C]//Proceedings of Neural Information Processing Systems. La Jolla: NIPS, 2021: 14745-14758.

[14] XIE E Z, WANG W H, YU Z D, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[C]//Proceedings of Neural Information Processing Systems. La Jolla: NIPS, 2021: 12077-12090.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021-01-03) [2022-06-22]. https://arxiv.org/abs/2010.11929.

[16] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 9992-10002.

[17] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 548-558.

[18] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. San Diego: JMLR, 2021: 7358-7367.

[19] LIN H Z, CHENG X, WU X Y, et al. CAT: Cross attention in vision transformer[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2022: 1-6.

[20] CHEN Z Y, ZHU Y S, ZHAO C Y, et al. DPT: Deformable patch-based transformer for visual recognition[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2899-2907.

[21] HAN K, XIAO A, WU E, et al. Transformer in transformer[C]//Proceedings of Neural Information Processing Systems. La Jolla: NIPS, 2021: 15908-15919.

[22] YUAN K, GUO S P, LIU Z W, et al. Incorporating convolution designs into visual transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 559-568.

[23] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1-9.

[24] TAN M X, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. San Diego: JMLR, 2019: 6105-6114.

[25] XIE S N, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 5987-5995.

[26] ZHU X Z, HU H, LIN S, et al. Deformable ConvNets V2: More deformable, better results[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9300-9308.

[27] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2022-06-22]. https://arxiv.org/

abs/2106.05786.
[28] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6230-6239.
[29] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.
[30] HUANG Z L, WANG X G, HUANG L C, et al. CCNet: Criss-cross attention for semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 603-612.
[31] CHEN C F R, FAN Q F, PANDA R. CrossViT: Cross-attention multi-scale vision transformer for image classification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 347-356.
[32] BA J L, KIROS J R, HINTON G E. Layer normalization [EB/OL]. (2016-07-21) [2022-06-22]. https://arxiv.org/abs/1607.06450.
[33] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.
[34] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6848-6856.
[35] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17) [2022-06-22]. https://arxiv.org/abs/1704.04861.
[36] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.
[37] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//European Conference on Computer Vision. Cham: Springer, 2014: 740-755.
[38] ZHOU B L, ZHAO H, PUIG X, et al. Semantic understanding of scenes through the ADE20K dataset[J]. International Journal of Computer Vision, 2019, 127(3): 302-321.
[39] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10425-10433.
[40] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]. (2019-01-04) [2022-06-22]. https://arxiv.org/abs/1711.05101.
[41] YUN S, HAN D, CHUN S, et al. CutMix: Regularization strategy to train strong classifiers with localizable features [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 6022-6031.
[42] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization[EB/OL]. (2018-04-27)[2022-06-22]. https://arxiv.org/abs/1710.09412.
[43] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2999-3007.
[44] CHEN K, WANG J Q, PANG J M, et al. MMDetection: Open MMLab detection toolbox and benchmark [EB/OL]. (2019-01-17)[2022-06-22]. https://arxiv.org/abs/1906.07155.

作者简介

胡　杰　男,1984年生,湖南永州人. 武汉理工大学汽车工程学院教授,博士生导师. 主要研究方向为汽车控制与诊断、车联网与大数据、智能驾驶、智能底盘等.
E-mail: auto_hj@163.com

昌敏杰　男,1999年生,湖北洪湖人. 武汉理工大学汽车工程学院硕士研究生. 主要研究方向为目标检测和目标跟踪.
E-mail: 1468139558@qq.com

徐博远　男,1998年生,湖北仙桃人. 武汉理工大学汽车工程学院硕士研究生. 主要研究方向为目标检测.
E-mail: 1903086417@qq.com

徐文才　男,1995年生,山东潍坊人. 武汉理工大学汽车工程学院博士研究生. 主要研究方向为3D目标检测、目标跟踪和场景理解.
E-mail: wencaixu_val@163.com