计算机科学
COMPUTER SCIENCE

基于Depth-wise卷积和视觉Transformer的图像分类模型

张峰, 黄仕鑫, 花强, 董春茹

引用本文

张峰, 黄仕鑫, 花强, 董春茹. 基于Depth-wise卷积和视觉Transformer的图像分类模型[J]. 计算机科学, 2024, 51(2): 196-204.

ZHANG Feng, HUANG Shixin, HUA Qiang, DONG Chunru. Novel Image Classification Model Based on Depth-wise Convolution Neural Network andVisual Transformer [J]. Computer Science, 2024, 51(2): 196-204.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100234

基于 Depth-wise 卷积和视觉 Transformer 的图像分类模型

张 峰 黄仕鑫 花 强 董春茹
河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室 河北 保定 071002
(fengzhang@hbu.edu.cn)

摘 要 图像分类作为一种常见的视觉识别任务,有着广阔的应用场景。在处理图像分类问题时,传统的方法通常使用卷积神经网络,然而,卷积网络的感受野有限,难以建模图像的全局关系表示,导致分类精度低,难以处理复杂多样的图像数据。为了对全局关系进行建模,一些研究者将 Transformer 应用于图像分类任务,但为了满足 Transformer 的序列化和并行化要求,需要将图像分割成大小相等、互不重叠的图像块,破坏了相邻图像数据块之间的局部信息。此外,由于 Transformer 具有较少的先验知识,模型往往需要在大规模数据集上进行预训练,因此计算复杂度较高。为了同时建模图像相邻块之间的局部信息并充分利用图像的全局信息,提出了一种基于 Depth-wise 卷积的视觉 Transformer(Efficient Pyramid Vision Transformer,EPVT)模型。EPVT 模型可以实现以较低的计算成本提取相邻图像块之间的局部和全局信息。EPVT 模型主要包含 3 个关键组件:局部感知模块(Local Perceptron Module,LPM)、空间信息融合模块(Spatial Information Fusion,SIF)和"+卷积前馈神经网络(Convolution Feed-forward Network,CFFN)。LPM 模块用于捕获图像的局部相关性;SIF 模块用于融合相邻图像块之间的局部信息,并利用不同图像块之间的远距离依赖关系,提升模型的特征表达能力,使模型学习到输出特征在不同维度下的语义信息;CFFN 模块用于编码位置信息和重塑张量。在图像分类数据集 ImageNet-1K 上,所提模型优于现有的同等规模的视觉 Transformer 分类模型,取得了 82.6%的分类准确度,证明了该模型在大规模数据集上具有竞争力。

关键词:深度学习;图像分类;Depth-wise 卷积;视觉 Transformer;注意力机制
中图分类号 TP391

Novel Image Classification Model Based on Depth-wise Convolution Neural Network and Visual Transformer

ZHANG Feng,HUANG Shixin,HUA Qiang and DONG Chunru
Hebei Key Laboratory of Machine Learning and Computational Intelligence,College of Mathematics and Information Science,Hebei University,Baoding,Hebei 071002,China

Abstract Deep learning-based image classification models have been successfully applied in various scenarios. The current image classification models can be categorized into two classes:the CNN-based classifiers and the Transformer-based classifiers. Due to its limited receptive field,the CNN-based classifiers cannot model the global relation of image,which decreases the classification accuracy. While the Transformer-based classifiers usually segmente the image into non-overlapping image patches with equal size,which harms the local information between each pair of adjacent image patches. Additionally,the Transformer-based classification models often require pre-training on large datasets,resulting in high computational costs. To tackle these problems,an efficient pyramid vision Transformer(EPVT) based on depth-wise convolution is proposed in this paper to extract both the local and global information between adjacent image patches at a low computational cost. The EPVT model consists of three key components:local perception module(LP),spatial information fusion module(SIF) and convolutional feed-forward network module(CFFN). The LP module is used to capture the local correlation of image patches. SIF module is used to fuse local information between adjacent image patches and improve the feature expression ability of the proposed EPVT by utilizing the long-distance dependence between different image patches. CFFN module is used to encode the location information and reconstruct tensors between feature image patches. To validate the proposed EPVT model's performance,various experiments are conducted on the benchmark data-

到稿日期:2022-11-28 返修日期:2023-06-14
基金项目:科技部重点研发项目(2022YFE0196100);河北省自然科学基金面上项目(F2018201115);河北省教育厅科学技术研究重点项目(ZD2019021);河北大学高层次创新人才科研启动经费项目
This work was supported by the National Key R&D Program of China(2022YFE0196100),Natural Science Foundation of Hebei Province(F2018201115),Key Scientific Research Foundation of Education Department of Hebei Province(ZD2019021) and Hebei University High-level Innovative Talent Research Start-up Funding Project.
通信作者:董春茹(dongcr@hbu.edu.cn)

sets，and experimental results show the EPVT achieves 82.6% classification accuracy on ImageNet-1K，which outperforms most of the SOTA models with lower computational complexity.

Keywords　Deep learning，Image classification，Depth-wise convolution，Visual transformer，Self-attention mechanism

1　引言

图像分类算法具有广泛的应用前景，目前已在人脸识别[1]、车辆违章检测[2]以及安防保障等现实场景中获得广泛的应用。在图像分类任务中，通常存在两个挑战。第一个挑战是冗余信息的处理。由于图像的局部区域通常包含趋于相近的冗余信息，而冗余信息容易引起低效的计算。早期关于图像冗余信息的处理方法大多是基于卷积神经网络的。最早在 2012 年，由 Krizhevsky 等[3]提出的 AlexNet 利用卷积神经网络(CNN)解决了图像中冗余信息处理的视觉问题。随着卷积神经网络层数的加深，出现了梯度消失的问题，为解决该问题，He 等[4]通过向每两层的卷积块中添加快捷连接，缓解了模型的梯度消失，使网络具有更好的特征提取能力。此后，基于卷积的图像分类网络[5-6]相继问世，并在计算机视觉领域占据了主导地位，成为了众多视觉任务的首选设计范式。然而，卷积运算虽然解决了输入特征图的局部冗余信息的问题，避免了不必要的计算开销，但是由于感受野有限，卷积网络难以学习图像的全局表示关系，而全局表示通常在视觉识别任务中起着至关重要的作用。为了解决这个问题，研究者试图扩展网络的深度和宽度，使用更大的卷积核[7]等方法，使 CNN 能够获得更大的感受野，但这也带来了新的问题，即网络参数的急剧增加。

第二个挑战是如何处理图像语义信息。图像通常具有复杂的语义信息，不同区域的目标之间通常存在着远距离依赖关系，远距离目标之间的信息交互往往会导致学习效率低下。受基于注意力的 Transformer 在自然语言处理领域的成功启发，许多研究者尝试探索 Transformer 架构在视觉领域的应用，ViT[8]开启了视觉领域一个新的里程碑，自此，研究者提出了许多强大的视觉 Transformer。多头注意力(Multi-Head Self Attention，MSA)是视觉 Transformer 的关键组件，MSA 模块采用输入特征之间的加权平均操作，通过计算输入特征上下文之间的相似度，动态地计算注意力权重[9]，这种方法允许注意力模块学习更多的特征，使 Transformer 能够捕获输入序列的远程依赖关系。不幸的是，Transformer 常常无法从网络的浅层中提取图像中细粒度的局部特征，从而降低了辨识图像背景和前景的能力。

综上，无论是卷积神经网络还是视觉 Transformer，都无法同时解决上述两个挑战。因此，如何学习图像局部特征和全局表示的问题仍然存在。为了同时应对上述的两个挑战，本文提出了一种基于 Depth-wise 卷积[10](Depth-wise Convolution，DW-Conv)和多头注意力的视觉 Transformer(EPVT)模型，该模型利用局部感知单元、空间信息融合模块和卷积前馈神经网络联合捕获图像的全局表示，提取相邻图像块的边缘和角等局部信息。本文的主要贡献概括如下：

1)将 Depth-wise 卷积组成的局部感知模块和 Transformer 中的编码器结合，提出了一种 EPVT 图像分类模型。该模型以少量的计算开销为代价，提高了对图像数据的全局表示和局部信息的建模能力，比单独利用注意力的模型或者精心设计的基于卷积和注意力的模型具有更优的分类性能。

2)将局部感知单元引入前馈神经网络，提出了具有 Depth-wise 卷积的前馈神经网络，该网络能够使用局部感知单元来额外学习特征图的 2D 表征信息。

3)提出了具有 Depth-wise 卷积的空间信息融合模块，并通过大量的消融实验验证了上述两个模型的有效性。在公开的图像分类数据集 ImageNet-1K 上，该模型优于同等参数量的模型，并取得了 82.6%的分类精度。

2　相关工作

2.1　基于卷积神经网络的图像分类模型

在视觉识别领域，卷积神经网络自出现以来一直被作为图像分类任务的主流架构。自从 LeCun 等提出了第一个标准的 CNN[11]以来，此后几十年，计算机视觉领域见证了各种各样的基于卷积神经网络在 ImageNet[12]数据集上取得前所未有的成功。ResNet 利用残差的概念实现了网络层数的加深，GoogleNet，DPN 和 MixNet 等进一步证明了基本块内多条路径的卷积运算的有效性。除了上述的架构进步以外，还有一些工作设计了高效的压缩算子，并在参数和精度方面做出了很好的平衡，使得网络能够适配移动端设备。

卷积神经网络采用局部建模设计的原则减小了需要学习的参数规模，更深的网络、更复杂的连接方式[13]、更大的规模[14]以及更复杂的卷积形式，使得网络具有更强大的特征提取能力。然而，传统的卷积具有固定的几何结构，无法有效地对不同尺度或者变形的目标进行建模。为了克服这一困难，经典的深度学习算法通过图像特征金字塔结构来生成不同尺度目标的特征表示，提升了模型应对复杂视觉任务的能力。本文将利用图像特征金字塔[15]结构的思想，对输入数据进行多尺度建模，并利用 Depth-wise 卷积和多头自注意力模块将图像中不同尺度的上下文信息编码到每个输出特征进行标记，学习图像中不同尺度目标的局部和全局的语义特征，提升模型的特征提取能力。

2.2　基于视觉 Transformer 的图像分类方法

自从 Dosovitskiy 等[8]提出 ViT 并在图像分类任务中取得巨大成功以来，一些工作尝试对视觉 Transformer 进行改进。为了降低 ViT 的训练成本，DeiT 提出了新的知识蒸馏，在不引入额外训练数据的情况下，使模型能够在 ImageNet-1k 数据集上展现出高性能。为降低全局自注意力引起的系统开销并使得相邻窗口的信息进行交互，Swin Transformer[16]采用了分层结构并提出窗口自注意力和 shift windows，该工作将自注意力全局建模的能力限制在窗口内，有效地减少了注意力模块引起的系统开销，shift windows 的设计方案使相邻窗口的信息得到了交互，提升了模型对局部信息的捕捉能力。此外，为了应对复杂的视觉任务，PVT[17]将

金字塔结构引入视觉 Transformer，使得模型能够生成不同尺度的特征表示，同时降低高像素输入引起的计算开销，以应对复杂的像素级和区域级的视觉任务。TNT[18]在 ViT 的基础上进行改进，解决了建模过程中 Patch 内部信息缺失的问题，使模型学习到更多块内部的局部信息。Zhou 等[19]在 Swin Transformer 的基础上，进一步分析了局部注意力模型表现平庸的原因，提出了一种增强局部自注意力模型 ELSA，改善了模型在多项视觉任务中的表现。除了以上的研究工作，研究者还致力于探索卷积和 Transformer 架构之间的兼容性。然而，基于 Transformer 的模型常常无法从网络的浅层中提取图像中细粒度的局部特征，从而降低了辨识图像背景和前景的能力。

2.3 基于 CNN 和 ViT 结合的图像分类算法

最近，一些工作使用 CNN 和 Transformer 共同建模的方法，试图解决上述问题。早期的工作通过大量的实验证明了卷积操作和 Transformer 架构具有互补性。例如，SENet 和 CBAM[20]的工作表明注意力模块可以作为卷积模块的增强；Conformer 采用 CNN 和 Transformer 的并行结构，提出特征耦合模块 FCU，对每个阶段的局部特征和全局特征进行特征对齐和信息交互；BoTNet[21]将注意力作为独立的组块来替代 CNN[22]模型中的传统卷积。另一些研究侧重于将注意力模块和卷积结合在单个块中（如 AA-ResNet），但该体系结构受限于为每个模块设计独立的路径。其他研究工作则侧重于将 MSA 和卷积结合在同一个 Block 中，如 ViATE[23]。此外，一些工作专注于向 Transformer[24]架构中引入卷积结构，使卷积和注意力模块在 Transformer[25-27]内部相互补充。尽管上述的工作在设计理念方面存在不同，但目的都是探索两者架构优势互补的可能性，并尝试为众多的视觉任务设计通用的视觉主干。在目前的工作中，大部分模型需要依赖一个预训练的 CNN 作为教师网络或使用 CNN 并行地参与训练，因此往往需要付出额外的训练成本，尤其是当卷积和注意力模块不能正确结合时，可能导致模型对局部特征表达能力的急剧恶化。

为解决上述两个挑战，本文提出了一种基于 Depth-wise 卷积的视觉 Transformer 图像分类模型。与标准的卷积核相比，Depth-wise 卷积的独特之处在于卷积核中的通道数始终为 1。在卷积运算过程中，特征图的每一个通道只使用一个卷积核，因此使用 Depth-wise 卷积核进行卷积运算后得到的特征图通道数量不变，并且在卷积运算中可产生更少的参数。

CNN 中卷积核的大小通常会影响模型的学习及训练。尽管使用大的卷积核，如 7×7 和 9×9 卷积，可以帮助模型获得更大的感受野，以覆盖特征图中更广泛的局部区域，但其缺点是会增加网络参数、计算量以及训练时间。使用多个 1×1 和 3×3 等小卷积来替代大卷积核是一种常见的策略，这种策略的优势体现在不仅可以减少计算量和参数，增加神经网络的深度，而且可以在网络中更多地使用非线性激活层，最终提升模型的语义判别能力。

为了平衡网络参数和模型的性能，本文通过消融实验选择在 EPVT 模型中使用 3×3 卷积作为卷积核的大小。EPVT 模型结构包含 3 个关键组件：局部感知模块、空间信息融合以及卷积前馈神经网络。其中局部感知单元模块用于捕获图像的局部相关性；空间信息融合模块用于融合相邻图像块之间的局部信息，并利用不同图像块之间的远距离依赖关系，提升模型的特征表达能力，使模型学习到输出特征在不同维度下的语义信息；卷积前馈神经网络模块则用于编码位置信息和重塑张量。

3 基于 Depth-wise 卷积和多头注意力的图像分类模型

图 1 给出了本文所提出的 EPVT 模型的总体架构，模型分为 4 个相似的阶段，每个阶段都包含一个图像块嵌入（Patch Embedding，PE）操作以及多个 Encoder 编码器。

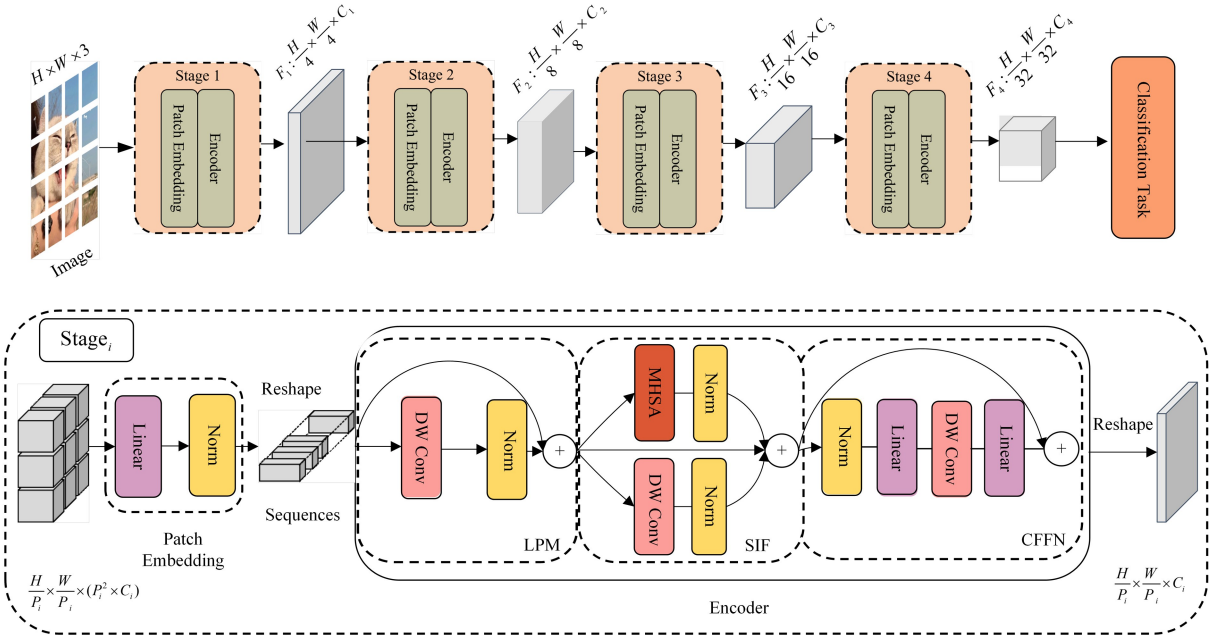

图 1 EPVT 模型的整体架构图

Fig. 1 Overall architecture of EPVT model

Encoder 编码器主要包含 3 个部分，即局部感知模块(Local Perceptron Module，LPM)、空间信息融合(Spatial Information Fusion，SIF)模块以及卷积前馈神经网络(Convolution Feed-Forward Network，CFFN)。其中，H 和 W 表示输入图像的高和宽，F_1，F_2，F_3 和 F_4 分别表示 4 个不同尺度的特征图，C_1，C_2，C_3 和 C_4 分别表示 4 个特征图的通道数；图像的下采样率分别是 4，2，2 和 2；P_i 表示第 i 阶段的采样率，i 分别取 1，2，3 和 4；“DW-Conv”表示 Depth-wise 卷积，它的分组数等于其通道数；MHSA(Multi-Head Self-Attention)表示多头注意力。

3.1　局部感知模块(LPM)

数据增强是视觉识别任务中常见的数据扩充技术，其中旋转和平移是常用的两种数据增强方法。由于 Transformer 包含较少的归纳偏置，通常依赖大规模的数据才能展现出较高的泛化性能。受文献[28]的启发，本文将使用 LPM 来增强模型的平移等方差性，并提升 Transformer 的建模能力。

如图 2 所示，LPM 由一个包含 3×3 的深度可分离卷积的残差结构块构成。输入特征序列 $X=[x_1,x_2,\cdots,x_n]$，$x_n\in R^D$，其中 n 是序列的长度，D 表示序列中每个字段的维度。首先输入序列 X，将其转化为具有 2 维结构的特征图 M，其次 LPM 模块将特征图 M 分别传导到两条并行的分支，包含 DW-Conv 卷积的分支将对特征图 M 进行卷积计算，而另一个分支则是对特征图的数据进行复制，然后将两个分支输出相加，生成新的特征图 M'，最后生成的特征图将被转化为新的序列 $\hat{X}$，并输入下一个模块。LPM 模块的计算式如式(1)—式(3)所示：

$$M=Reshape(X) \tag{1}$$

$$M'=Norm(DWConv(M))+M \tag{2}$$

$$\hat{X}=Flatten(M') \tag{3}$$

其中，$DWConv(\cdot)$ 表示 depth-wise 卷积操作；$Norm$ 表示层标准化；$Flatten(\cdot)$ 表示特征映射函数，作用是将特征图 M' 映射成为特征序列 $\hat{X}$。

LPM 采用 Depth-wise 卷积，卷积核大小为 3×3，步长设为 2，卷积核每次沿着一个方向进行滑动建模时，卷积核扫描到的两个相邻区域会部分重叠，通过该方式能够学习到相邻图像块之间的空间信息，使模型能够学习到更多的局部特征。

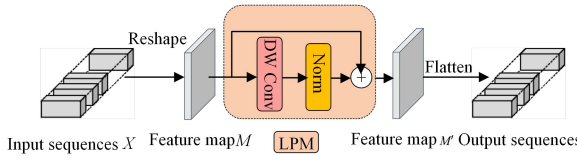

图 2　局部感知模块结构图

Fig. 2　Diagram of LPM module structure

3.2　空间信息融合模块

空间信息融合模块由一个 3×3 的 Depth-wise 卷积[29]和一个多头自注意力模块构成，每个 SIF 中都包含残差结构(Residual Module，RM)[30]和层次标准化(Layer Normalization，LN)[31]。

如图 3 所示，SIF 模块将输入序列 X_{input} 分别输送到包含 MSA、DWConv 和残差的 3 个并行分支中。在包含 DWConv 的分支中，首先，X_{input} 将转换为二维的特征图并经过 DW-Conv 进行卷积运算，紧接着进行层标准化，最后进行拉平并转换为新的序列 X_{dwc}。在残差分支中，序列 X_{input} 保持原有的算术值；在包含 MSA 的分支中，首先序列 X_{input} 将输入到 MSA 模块中并完成对序列中各个向量间的全局建模，其次经过一个层标准化，最后生成一个新的序列 X_{msa}。3 个并行分支的输出 X_{dwc}，X_{msa} 和 X_{input} 将进行求和运算，并最终生成输出序列 X_{output}。

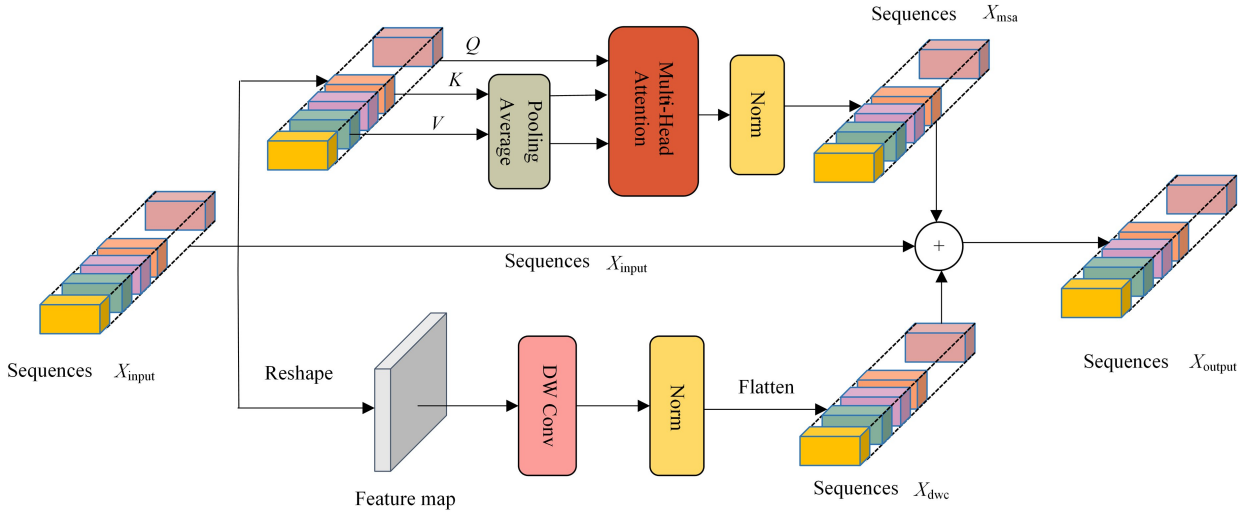

图 3　空间信息融合模块和线性空间缩减注意力模块

Fig. 3　SIF module and linear SRA module

为降低模型的计算复杂度，本文采用文献[17]提出的具有线性复杂度的空间自注意力(Linear Spatial Reduction Attention，LSRA)，如图 3 所示。与传统的 MSA 结构类似，LSRA 仍然采用查询 Query(Q)、键 Key(K)、值 Value(V)作为输入，计算过程见式(4)、式(5)：

$$LSRA(Q,K,V)=Concat(head_1,\cdots,head_{N_i})w^o \tag{4}$$

$$head_j=Att(Qw_j^Q,AvgP(K)w_j^K,AvgP(V)w_j^V) \tag{5}$$

其中，N_j 表示第 i 阶段注意力层的多头个数，$w_j^Q\in\mathbb{R}^{C_i\times d_{head}}$，

$w_j^K \in^{C_i \times d_{\text{head}}}$，$w_j^V \in^{C_i \times d_{\text{head}}}$ 分别表示第 i 阶段的线性投影，Avg$P(\cdot)$表示平均池化，目的是减少参数运算量。$Att(\cdot)$的计算式如下：

$$Att(Q,K,V)=\text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_{\text{head}}}}\right)V \tag{6}$$

LSRA 像卷积层一样享有线性计算的内存成本，具体来说，给定大小为 $h\times w\times c$ 的输入，LSRA 的时间复杂度为：

$$\Omega(LinearSRA)=2hwP^2c$$

其中，P 代表池化的大小。

3.3 卷积前馈神经网络(CFFN)

在 Transformer 中，前馈神经网络(Feedfor-ward Neural Network，FNN)是一种基础的特征增强模块，通常由一个或者多个线性变换的全连接层和非线性激活函数组成。传统的前馈神经网络仅用数个全连接层在特征的通道上进行增强，作用是将输入的词向量经过一系列线性变换和激活函数处理后输出另一个词向量，并没有考虑到图像是二维特征。因此，我们在传统的 FFN 模块中添加了一个 3×3 的 Depth-wise 卷积，用于额外地学习图像的二维特征。

受文献[17]的启发，本文不再使用固定大小的位置编码，并采用了与文献[17]相同的零填充(Zero Padding，ZP)位置编码，如图 4 所示。输入序列 X_{input} 将分别传送到两个分支中。

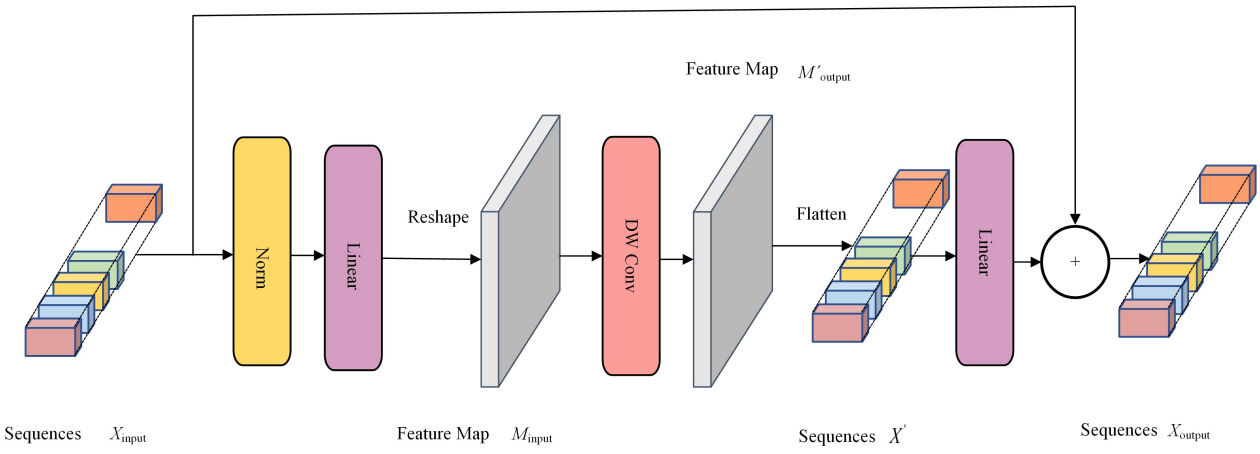

图 4 前馈神经网络模块结构图

Fig. 4 Diagram of CFFN module

EPVT 算法如算法 1 所示。

算法 1 EPVT 算法

输入：网络 EPVT=(B，N_{epochs}，X_{labels})；EPVT 模型中 4 个阶段的堆叠次数 b_i，i=1，2，3，4

输出：预测概率

```
1. /* 训练一个序列模型 */
2. 初始化 EPVT
3. for epoch←1 to N_epochs do
4.     for i←1 to n do
5.         if i≤1 then
6.             使用采样率为 4 的下采样函数对输入数据进行下采样，
               并获得特征图 M；
7.             else
8.                 使用采样率为 2 的下采样函数对输入数据进行下采
                   样，并获得特征图 M；
9.             使用线性映射函数将特征图 M 映射成特征序列 X；
10.            for b←b_1 to b_4 do
11.                X'←LPM(X)；
12.                X_sif←SIF(X')；
13.                X_effn←CFFN(X_sif)；
14.        end for
15.    end for
16. X_output←Flatten(X_effn)；
17. 计算预测概率，P=loss(X_output，X_labels)；
18. end for
19. return P
```

算法 1 中，步骤 13 对应的计算式如下：

$$M_{\text{input}}=Reshape(Linear(Norm(X_{\text{input}})) \tag{7}$$

其中，X_{input} 表示输入序列，Norm 表示 Layer 标准化，Linear 表示全连接层，Reshape 表示将序列转化为二维图像的运算操作，M_{input} 表示生成的特征图。

$$M'_{\text{output}}=DWConv(M_{\text{input}}) \tag{8}$$

其中，DWConv 表示深度可分离卷积运算，M'_{output} 表示 M_{input} 经过 Depth-wise 卷积后得到的特征图。

$$X'=Flatten(M'_{\text{output}}) \tag{9}$$

其中，$Flatten$ 表示对二维的特征图的序列化操作，X' 表示由特征图 M'_{output} 经过 $Flatten$ 后得到的序列。

$$X_{\text{output}}=X_{\text{input}}+Linear(X') \tag{10}$$

其中，$Linear$ 表示全连接层，X_{output} 表示 X_{input} 和经过全连接层的 X' 生成的序列的总和。

4 实验与分析

为验证本文提出的 EPVT 图像分类模型的性能，我们分别在 ImageNet-1K 和 ImageNet-100 上进行了大量的比较实验和消融实验，实验结果证明，模型在图像分类数据集 ImageNet-1K 上优于现有的同等规模的视觉 Transformer 分类模型。

4.1 实验环境设置

4.1.1 数据集处理

本文选取的数据集来源于公开的大规模图像数据

ImageNet-1K。在 2009 年，该图像数据在一项研究中被推出，它是模型大规模图像识别任务的基准数据集。本文实验的数据集如下。

1) ImageNet-1K：本实验选取 ILSVRC 竞赛中常用的 ILSVRC-2012 作为第一个数据集，它包含 128 万张训练示例以及 5 万张验证图像。

2) ImageNet-100：ImageNet-100 数据集作为 ImageNet-1K 的子集，包含 10 万张训练数据以及 2 万张验证数据，每一类数据包含 50 张测试样例。

4.1.2　基准方法

本小节将 EPVT 模型与以下提到的几类主流的基准方法进行比较，它们分别是基于卷积神经网络的图像分类模型，如 ResNet，EfficientNet，MoblieNetV2；基于 Transformer 的分类模型，如 Swin Transformer，PVTv2，LocalViT，T2T-ViT，TNT，CrossViT；基于卷积和 Transformer 的混合模型，如 ViTAE，Conformer，DeiT。

4.1.3　模型参数设置

本文的实验设备为 2 张 NVIDIA RTX 3090 GPU，并基于 Pytorch 1.10.1 框架，使用 Python3.8 来实现 EPVT 模型。该模型主要由 4 个阶段组成，首先模型将对输入特征图进行下采样，经过下采样的特征图的尺寸分别为 56×56，28×28，14×14 和 7×7。在第 2 章提及的局部感知单元、空间信息融合模块和卷积前馈神经网络中，实验默认使用窗口大小为 3 的 Depth-wise 卷积，每个卷积核的参数由随机初始化获得。在 EPVT 模型中，本文将 MLP 的扩张率设置为 4，除作为最终预测组件的全连接层采用 GeLU 激活函数外，其余的 MLP 均采用 ReLU[32] 激活函数。

表 1　EPVT 架构参数

Table 1　Parameters of EPVT architecture

	Output Size	Layer Name	EPVT		
			Tiny	Small	Large
1	$\frac{H}{4}\times\frac{W}{4}$	Overlapping Patch Embedding	$S_1=4$		
			$C_1=32$	$C_1=64$	$C_1=64$
		Transformer Encoder	$R_1=8$	$R_1=8$	$R_1=8$
			$H_1=1$	$H_1=1$	$H_1=1$
			$E_1=8$	$E_1=4$	$E_1=8$
			$B_1=2$	$B_1=2$	$B_1=3$
2	$\frac{H}{8}\times\frac{W}{8}$	Overlapping Patch Embedding	$S_2=2$		
			$C_2=64$	$C_2=128$	$C_2=128$
		Transformer Encoder	$R_2=4$	$R_2=4$	$R_2=4$
			$H_2=2$	$H_2=2$	$H_2=2$
			$E_2=8$	$E_2=8$	$E_2=8$
			$B_2=2$	$B_2=2$	$B_2=4$
3	$\frac{H}{16}\times\frac{W}{16}$	Overlapping Patch Embedding	$S_3=2$		
			$C_3=160$	$C_3=320$	$C_3=320$
		Transformer Encoder	$R_3=2$	$R_3=2$	$R_3=2$
			$H_3=5$	$H_3=5$	$H_3=4$
			$E_3=4$	$E_3=4$	$E_3=4$
			$B_3=2$	$B_3=6$	$B_3=6$
4	$\frac{H}{32}\times\frac{W}{32}$	Overlapping Patch Embedding	$S_4=2$		
			$C_4=256$	$C_4=512$	$C_4=512$
		Transformer Encoder	$R_4=1$	$R_4=1$	$R_4=1$
			$H_4=8$	$H_4=8$	$H_4=8$
			$E_4=4$	$E_4=4$	$E_4=4$
			$B_4=2$	$B_4=2$	$B_4=3$

EPVT 模型架构的详细超参数如表 1 所列。令 $i(i=1,2,3,4)$ 表示 EPVT 模型的第 i 个阶段，则 S_i 表示第 i 个阶段特征图的下采样率；C_i 表示第 i 个阶段特征图的输出通道数；H_i 表示第 i 个阶段使用 H 个多头注意力；R_i 表示第 i 个阶段的 SRA 的减少率；E_i 表示在第 i 个阶段的前馈网络中，隐藏层相对于输入层的膨胀率；B_i 表示第 i 个阶段 Transformer 编码器的堆叠次数。

表 2　实验参数设置

Table 2　Experimental parameter settings

Train Parameter	Value
自动增强	rand-m9-mstd0.5-incl
颜色抖动	0.4
随机裁剪	1.0
混合标签	0.8
混合概率	1.0
基础学习率	0.0005
总迭代次数	300
优化器 EPS	1×10^{-8}
优化器 BETAS	(0.9,0.999)
优化器动量参数	0.9
预热轮数	20
预热学习率	5×10^{-7}
权重衰减	0.05

如表 1 所列，EPVT 模型中包含 4 个部分重叠的块嵌入(Overlapping Patch Embedding，OPE)，OPE 的作用是对输入特征图进行下采样，并将其转换为序列化的特征向量，最后输入 Transformer 编码器中。

OPE 的采样过程如图 5 所示，图中深色的圆点表示输入特征图；浅色的圆点表示对输入特征图进行填充，填充值为 0。

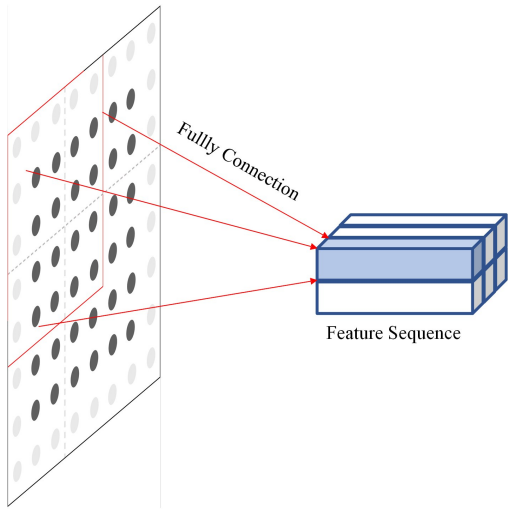

图 5　部分重叠的块嵌入

Fig. 5　Overlapping patch embedding

实验的数据增强的超参数如表 2 所列。其中优化器 EPS 默认值为 1×10^{-8}，它的作用体现在可以提高梯度反传数值的稳定性。优化器 BETAS 用于计算梯度的运行的平均值及其平方的系数。自动增强、颜色抖动以及随机裁剪的作用体现在对原始的输入图像进行预处理。

表 3 ImageNet-1K 数据集上 EPVT 和 SOTA 方法的比较

Table 3 Comparison of EPVT and SOTA methods on ImageNet-1K dataset

Arch.	Model	Params	FLOPs	Input size	ImageNet Top-1/%
基于 CNN 的分类模型	ResNet-18	11.7×10^6	3.6×10^9	224	70.3
	ResNet-50	25.6×10^6	7.6×10^9	224	76.7
	ResNet-101	44.5×10^6	15.2×10^9	224	78.3
	ResNet-152	60.2×10^6	22.6×10^9	224	78.9
	EfficientNet-B0	5.3×10^6	0.8×10^9	224	77.1
	EfficientNet-B4	19.3×10^6	8.4×10^9	224	82.9
	MobileNetV1	4.3×10^6	0.6×10^9	224	72.3
	MobileNetV2(1.4)	6.9×10^6	0.6×10^9	224	74.7
基于 Transformer 的分类模型	DeiT-T	5.7×10^6	2.6×10^9	224	72.2
	PVTV2-B0	3.4×10^6	0.6×10^9	224	70.5
	LocalViT-T2T	4.3×10^6	2.4×10^9	224	72.5
	T2T-ViT-7	4.3×10^6	1.2×10^9	224	71.7
	EPVT-T	$\mathbf{3.4\times10^6}$	$\mathbf{0.6\times10^9}$	224	**73.0**
	CrossViT-Ti	6.9×10^6	3.2×10^9	224	73.4
	ViTAE-6M	6.5×10^6	2.0×10^9	224	77.9
	PVTV2-b1	13.1×10^6	2.1×10^9	224	78.7
	LocalViT-PVT	13.5×10^6	9.6×10^9	224	78.2
	EPVT-S	14.1×10^6	2.1×10^9	224	**79.6**
	DeiT-S	22.1×10^6	9.8×10^9	224	79.9
	PVTV2-b2-li	22.6×10^6	3.9×10^9	224	82.1
	Conformer-Ti	23.5×10^6	5.2×10^9	224	81.3
	Swin-T	29.0×10^6	9.0×10^9	224	81.3
	TNT-S	23.8×10^6	10.4×10^9	224	81.4
	ViTAE-S	23.6×10^6	5.6×10^9	224	82.0
	EPVT-L	22.3×10^6	$\mathbf{3.8\times10^9}$	224	**82.6**

4.2 对比实验结果

表 3 列出了各个基准模型和 EPVT 在 ImageNet-1K 上的性能表现，其中参数量的单位为“百万”。表 3 中，相比其他基准算法，在该数据集上，本文提出的模型具有较高的分类精度。与 PVT v2 和 Swin Transformer 等纯视觉 Transformer 相比，EPVT 模型在精度方面具有很大的优势，如 EPVT-L 实现了 82.6%的 ImageNet Top-1 准确度。

此外，模型所需的 FLOPs 更少，说明本文模型的训练速度要优于两者。与 VITAE 和 Conformer 等近期基于卷积和 Transformer 的模型相比，本文模型在参数规模相近的情况下，具有更好的性能表现以及更少的每秒浮点运算次数(Floating-Point Operations Per Second，FLOPs)。

4.3 消融实验结果分析

本小节为了研究 EPVT block 内部各改进之处对模型性能的影响，在公开数据集 ImageNet-100 上分别对其进行消融实验。为了保证实验的公平性以及 EPVT 各个组件的有效性，本文选择 EPVT-T 作为基准模型，并采用相同的优化策略、数据增强方式以及超参数对模型的 3 个组件进行消融研究。此外，本小节的所有实验结果均经过 300 个 epoch 的迭代后获得。

为验证不同模块对 EPVT 模型性能的影响，我们分别做了 5 组实验，实验结果如表 6 所列。

实验 1 不使用 LPM 模块，并移除 SIF 和 CFFN 中的所有卷积组件，以观察模型的性能表现。

实验 2 仅在实验 1 的基础上添加 SIF 和 CFFN 模块并移除 LPM 模块，以验证 LPM 模块组件对模型性能的影响。

实验 3 仅在实验 1 的基础上对 SIF 模块中的旁路卷积分支进行消融，因此添加了 LPM 和 CFFN 模块，并移除 SIF 组件中的旁路分支，以验证该分支对模型分类性能的影响。

实验 4 仅在实验 1 的基础上对 CFFN 模块中的卷积组件进行消融，因此添加了 LPM 和 SIF 模块，并移除 CFFN 组件中的 Depth-wise 卷积模块，以验证该模块对模型分类性能的影响。

实验 5 为本文提出的完整 EPVT 模型，它包含第 2 章提及的 3 个模块，分别是 LPM，SIF 和 CFFN。

表 4 中，w/o DC 代表不包含 Depth-wise 卷积模块，#Param表示模型的参数量，Top-1 的单位为“%”，(w/o DC)表示不含 Depth-wise 卷积，Y 表示包含该模块，N 表示不包含该模块。实验数据如表 4 所列。

表 4 消融实验结果对比表

Table 4 Comparison of ablation experiment results

Experiment	LPM	SIF (w/o DWConv)	CFFN (w/o DWConv)	ImageNet-1000	
				#parameter	Top-1
实验 1	N	N	N	3.2×10^6	89.0(−1.3)
实验 2	Y	Y	N	3.2×10^6	89.8(−0.4)
实验 3	Y	N	Y	3.2×10^6	89.2(−1.0)
实验 4	N	Y	Y	3.2×10^6	87.3(−2.9)
实验 5	Y	N	N	3.2×10^6	90.2

4.3.1 局部感知模块对模型性能的影响分析

对比实验 1 和实验 5，当 EPVT-T 不使用 LPM 模块时，模型的精度由 90.2%下降到 89.0%。通过实验发现，LPM 模块对实验精度的影响最大，这表明更早的卷积能够大幅度地提升视觉 Transformer 的性能。此外，Xiao 等[28]的工作也对该项实验结论起到了支撑作用。

4.3.2 空间信息融合(w/o DWC)模块对实验精度的影响

SIF 模块包含了局部感知的 Depth-wise 卷积以及线性

计算复杂度的多头自注意力。对比表中实验 2 和实验 5 的结果可以发现，当 SIF 不包含 Depth-wise 卷积模块时，模型的精度由 90.2%下降至 89.8%，下降了 0.4 个百分点。实验表明，采用两条并行的数据处理分支，并将两者的数据输出进行融合，有利于增强视觉 Transformer 的图像预测能力。此外，通过实验 3 和实验 6 可以发现，SIF 模块包含 Depth-wise 卷积模块对模型精度的影响大于 CFFN 不包含 Depth-wise 卷积模块。

4.3.3　卷积前馈网络(w/o DWC)模块对实验精度的影响

对比实验 3 和实验 5，我们发现含 DWC 的 CFFN 模块对模型实验精度的影响仅次于 LPM 模块，并高于 CFFN。此外，缺少 DC 模块的 CFFN 组件使模型的最终分类结果下降，而这一结果是可预期的，原因是本文的 CFFN 结构中包含隐式的位置编码。视觉 Transformer 中的位置编码对图像语义信息的捕获是至关重要的，由于 Transformer 的位置无关性，通常需要借助位置编码来指导模型学习图像的原始语义，相关工作[8,27]的结论能够支撑这一观点。

4.3.4　卷积核的尺寸对模型性能的影响

本小节分别选取尺寸大小为 3×3，5×5 和 7×7 的卷积核进行实验验证，以研究不同尺寸卷积核对 EPVT 模型的精度的影响。如图 6(a)所示，随着卷积核尺寸的增大，EPVT-T 模型的精度呈现下降的趋势。造成这一结果的原因可能是大的卷积核，例如大于 3×3 的卷积，在图像特征提取的过程中引入了过多的冗余信息，造成了模型建模能力的局部恶化。

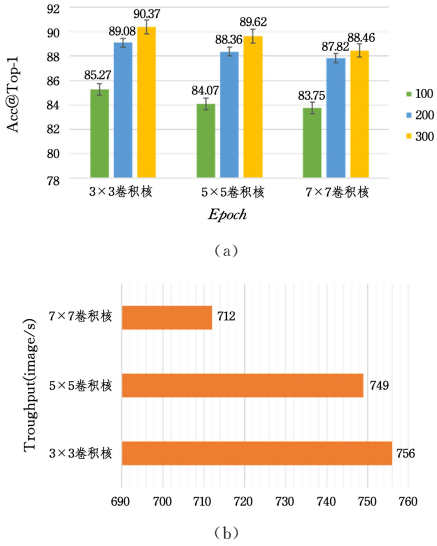

图 6　不同滤波器尺寸的 EPVT 分类精度和吞吐量

Fig. 6　EPVT classification accuracy and throughput with different filter sizes

此外，本节还研究了不同大小的卷积核对模型吞吐量的影响。吞吐量指单位时间内操作系统所处理的数据量，在神经网络中，吞吐量指网络在单位时间内可以处理的最大输入实例数。虽然吞吐量受到设备的存储速度、CPU 利用率、数据的并行化和网络带宽的影响，但吞吐量的大小主要取决于主存储器在两次启动操作之间需要的最小时间间隔。与处理单个实例的延迟不同，我们希望设备尽可能并行地处理多个实例，以获得最大的吞吐量。然而，并行性显然依赖于数据、模型和设备。因此，为了正确测试具有不同大小卷积核的 EPVT-T 模型的吞吐量，尽可能地避免存储设备、CPU 和数据并行化对结果的影响，我们执行以下两个步骤：1)设置最大并行实例数为 128；2)在相同的数据集和实验设备 ImageNet-100 和 NVIDIA RTX 3090 上，测量模型每秒可以处理的实例数。如图 6(b)所示，实验表明，当卷积核的大小为 3×3 时，模型的吞吐量达到 756，高于其他尺寸的 EPVT-T 模型。

结束语　针对基于卷积神经网络的模型无法对图像的全局关系进行建模，而基于 Transformer 的模型则通常需要对大型数据集进行预先训练，导致计算成本很高的问题，本文提出了一种基于深度卷积的高效金字塔视觉转换器(EPVT)，以较低的计算成本提取相邻图像块之间的局部和全局信息。为了验证所提出的 EPVT 模型的性能，在基准数据集上进行了对比实验和消融，实验结果表明 EPVT 在 ImageNet-1K 上的分类准确率达到 82.6%，优于大多数目前流行的视觉处理模型。

未来，我们主要在两个方面对模型进行改进。1)探索可变形卷积和视觉 Transformer 结合的可行性，其目的是借助可变形卷积处理不同尺度目标的优势，增强视觉 Transformer 应对复杂视觉任务的能力，例如目标检测和图像分割。2)本文将进一步探究每一个阶段的注意力组件对模型实验精度的影响，并思考对注意力组件进行优化，加快模型的训练速度并提升预测的准确度。

参 考 文 献

[1] ZHU Z, HUANG G, DENG J, et al. Webface260m: a benchmark unveiling the power of million-scale deep face recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2021: 10492-10502.

[2] LIU X, ZHANG P, YU C, et al. Watching you: Global-Guided reciprocal learning for video-based person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2021: 13334-13343.

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

[4] 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. Piscataway: IEEE Computer Society, 2016: 770-778.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 2014, 1409.1556.

[6] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 2017, 1704.04861.

[7] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv: 2015, 1511.07122.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recogni-

tion at scale[J]. arXiv:2020,2010.11929.
[9] ZHOU L Y,YUAN T T,CHEN S Y. Sequence-to-sequence sign language recognition and translation in Chinese continuous sign language [J]. Computer Science,2022,49(9):155-161.
[10] HU F Y,WANG X J,SHEN M F,,et al. Research progress of image instance segmentation by deep convolutional neural network [J]. Computer Science,2022,49(5):10-24.
[11] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.
[12] 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 Computer Society,2009:248-255.
[13] SZEGEDY C,LIU W,JIA Y,et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway:IEEE Computer Society,2015:1-9.
[14] TAN M,LE Q. Efficientnet:Rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. New York:PMLR,2019:6105-6114.
[15] BAY H,TUYTELAARS T,GOOL L V. Surf:Speeded up robust features[C]// European Conference on Computer Vision. Berlin:Springer,2006:404-417.
[16] LIU Z,LIN Y,CAO Y,et al. Swin transformer:Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE Computer Society,2021:10012-10022.
[17] WANG W,XIE E,LI X,et al. Pvt v2:Improved baselines with pyramid vision transformer[J]. Computational Visual Media,2022,8(3):415-424.
[18] HAN K,XIAO A,WU E,et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems,2021,34:15908-15919.
[19] ZHOU J,WANG P,WANG F,et al. ELSA:Enhanced local self-attention for vision transformer[J]. arXiv:2021,2112.12786.
[20] WOO S,PARK J,LEE J Y,et al. CBAM:Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision(ECCV). Berlin:Springer,2018:3-19.
[21] SRINIVAS A,LIN T Y,PARMAR N,et al. Bottleneck transformers for visual recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Computer Society,2021:16519-16529.
[22] CHEN C Q. Development of convolutional neural network and its application in computer vision [J]. Computer Science,2019,46(3):63-73.
[23] XU Y,ZHANG Q,ZHANG J,et al. Vitae:Vision transformer advanced by exploring intrinsic inductive bias[J]. Advances in Neural Information Processing Systems,2021,34:28522-28535.
[24] ZHANG J H,LIU F,QI J Y. A bottleneck transformer based lightweight micro-expression recognition architecture [J]. Computer Science,2022,49(6A):370-377.
[25] LI K,WANG Y,ZHANG J,et al. Uniformer:Unifying convolution and self-attention for visual recognition[J]. arXiv:2022,2201.09450.
[26] 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. Piscataway:IEEE Computer Society,2021:22-31.
[27] CHU X,TIAN Z,ZHANG B,et al. Conditional positional encodings for vision transformers[J]. arXiv:2021,2102.10882.
[28] XIAO T,SINGH M,MINTUN E,et al. Early convolutions help transformers see better[J]. Advances in Neural Information Processing Systems,2021,34:30392-30400.
[29] CHEN Y,DAI X,CHEN D,et al. Mobile-former:Bridging mobilenet and transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Computer Society,2022:5270-5279.
[30] WANG Y,YANG Y,BAI J,et al. Evolving attention with residual convolutions[C]// International Conference on Machine Learning. New York:ACM 2021:10971-10980.
[31] BA J L,KIROS J R,HINTON G E. Layer normalization [J]. arXiv:2016,1607.06450.
[32] GLOROT X,BORDES A,BENGIO Y. Deep sparse rectifier neural networks[C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Cambridge:JMLR Workshop and Conference Proceedings,2011:315-323.

ZHANG Feng, born in 1976, Ph.D, associate professor, master supervisor, is a member of CCF (No. 65203M). Her main research interests include machine learning and intelligent decision-making.

DONG Chunru, born in 1980, Ph.D, associate professor, master supervisor. His main research interests include deep learning and image processing.

(责任编辑:喻藜)