文章编号:1671-6833(2024)04-0019-11

面向图像分类的 Vision Transformer 研究综述

智 敏, 陆静芳

(内蒙古师范大学 计算机科学技术学院,内蒙古 呼和浩特 010022)

摘 要: 作为一种基于 Transformer 架构的模型,ViT 已经在图像分类任务中展现出了良好的效果。对 ViT 在图像分类任务上的应用进行系统性归纳总结。首先,简单介绍了 ViT 框架及其 4 个模块(patch 模块、位置编码、多头注意力和前馈神经网络)的功能特性;其次,以 ViT 中 4 个模块的改进措施为脉络综述其在图像分类任务中的应用;再次,由于不同的模型结构和改进措施对最终的分类性能产生显著影响,还对文中出现的各类 ViT 进行了横向对比,并详细列出模型的参数和分类精度及其优缺点;最后,指出 ViT 在图像分类任务中的优势和局限性,并提出未来可能的研究方向以打破其局限性,进一步扩展 ViT 在其他计算机视觉任务中的应用,同时,还可以探索将 ViT 扩展到视频理解等更广泛的计算机视觉领域。

关键词:ViT 模型; 图像分类; 多头注意力; 前馈网络层; 位置编码

中图分类号:TP181;TP391 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.01.015

卷积神经网络[1-3](convolution neural network, CNN)一直以来是计算机视觉领域的主导模型,但随着 Transformer[4] 在自然语言处理领域的广泛应用,将其扩展到视觉任务已经成为当今的研究热点之一。ViT(Vision Transformer)由 Dosovitskiy 等[5]提出,利用图像块作为模型的输入,应用于目标检测[6]、实例分割[7-8]、跟踪[9]、图像生成[10]和图像增强[11]等多项视觉任务中。

在此之前,许多研究者已经对 ViT 在图像分类任务中的应用进行了综述。Tay 等[12]回顾了 ViT 在语言任务中的应用;Khan 等[13]和 Han 等[14]总结了早期的 ViT 模型和基于注意力的模型;Lin 等[15]提供了关于多种面向视觉应用的最新 ViT 系统的综述。此外,还有一些从算法角度对图像任务进行的综述研究。毕莹等[16]对遗传算法在图像分析领域的代表性研究工作,如特征提取、图像分类、边缘检测和图像分割等进行了全面且系统的讨论和综述。与其他综述不同的是,本文的目的是回顾最新出现的 ViT 改进方法在图像分类任务中的应用,并对其进行系统分类。①对于多个基于 ViT 模型改进的图像分类应用,本文选择了代表性的方法进行全面回顾,并进行详细描述和分析。除了独立分析每个模型,还在一定程度上建立了它们之间的内部联系。②由于现有的 ViT 方法针对分类任务采用不同的训练方案和超参数设置,本文在不同的数据集和限制条件下进行了多次横向比较。更重要的是,本文总结了对 ViT 不同模块的改进,包括 patch 模块、位置编码、多头注意力和前馈神经网络,并进行了总结、分析和比较。③本文还阐述了不同模型的优缺点,并对实验结果进行了分析比较,并提出了 ViT 未来可能的研究方向。

1 ViT 框架与组成

ViT 整体网络结构如图 1 所示,是第一个用于图像分类的 Transformer 主干网络。ViT 继承了标准 Transformer 的编码器结构,主要包含 patch 模块、位置编码、多头注意力和前馈神经网络,以下将对这 4 部分内容进行具体介绍。

1.1 patch 模块

标准的 Transformer 接收一个一维的 token 嵌入序列作为输入。因此,在处理图像任务时,要将二维输入图像 $x \in \mathbf{R}^{H\times W\times C}$ 重构为一系列展平的图像块

收稿日期:2023-08-15;**修订日期**:2023-10-10

基金项目:内蒙古自治区自然科学基金资助项目(2023MS06009);内蒙古师范大学基本科研业务费专项基金项目(2022JBXC018);内蒙古师范大学研究生科研创新基金项目(CXJJS22138)

作者简介:智敏(1972—),女,内蒙古赤峰人,内蒙古师范大学教授,博士,主要从事深度学习、视频图像处理的研究,E-mail:cieczm@imnu.edu.cn。

引用本文:智敏,陆静芳. 面向图像分类的 Vision Transformer 研究综述[J]. 郑州大学学报(工学版),2024,45(4):19-29.(ZHI M, LU J F. A review of Vision Transformer for image classification[J]. Journal of Zhengzhou University (Engineering Science),2024,45(4):19-29.)

$x \in \mathbf{R}^{N\times(P^2\times C)}$,即不重叠的 patch,此过程通过卷积实现。其中,(H,W) 为原始图像的分辨率;(P,P) 为每个图像块的分辨率;$N = HW/P^2$ 为产生的图像块数,同时也是 ViT 的有效输入序列长度。ViT 在其所有层中需使用恒定的隐藏向量的大小 D,因此,要将图像块展平,并使用可训练的线性投影来将其映射到 D 维,将此投影的输出称为块嵌入(patch embedding)。

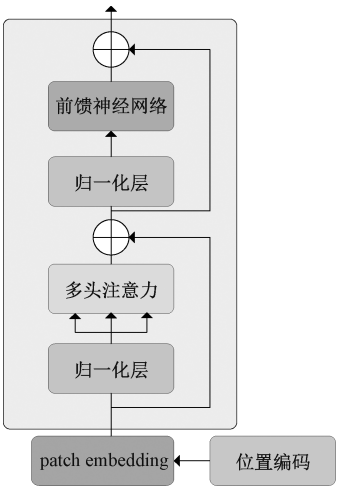


图 1 ViT 网络结构图

Figure 1 ViT network structure

1.2 位置编码

位置编码被添加到 patch embedding 层中以保留位置信息。ViT 使用标准的可学习一维位置嵌入。位置编码有很多选择,例如,不同频率的正弦和余弦函数如下所示:

$$\mathrm{PE}_{(pos,i)} = \begin{cases} \cos pos_{\omega_k}, i = 2k+1; \\ \sin pos_{\omega_k}, i = 2k。 \end{cases} \tag{1}$$

式中:$\omega = \dfrac{1}{10\ 000^{\frac{2k}{d}}}$;$k = 1,2,\cdots,\dfrac{d}{2}$;$i$ 和 d 分别为向量的索引和长度;pos 为序列中每个元素的位置。

1.3 多头注意力

如图 2 所示,多头注意力机制将输入线性投影到多个特征子空间中,并通过几个独立的注意力头并行处理,然后向量被并联映射到最终的输出。多头注意力机制的过程可表述为

$$\begin{cases} \boldsymbol{Q}_i = \boldsymbol{X}\boldsymbol{W}^{Q_i}; \\ \boldsymbol{K}_i = \boldsymbol{X}\boldsymbol{W}^{K_i}; \\ \boldsymbol{V}_i = \boldsymbol{X}\boldsymbol{W}^{V_i}。 \end{cases} \tag{2}$$

$$\boldsymbol{Z}_i = \mathrm{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)\ ,i = 1,2,\cdots,h; \tag{3}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_h)\ \boldsymbol{W}^O。 \tag{4}$$

式中:h 为头的数量;$\boldsymbol{W}^O \in \mathbf{R}^{hd_v\times d_{\mathrm{model}}}$ 为输出映射矩阵;$\boldsymbol{Z}_i$ 为每个头的输出向量;$\boldsymbol{W}^{Q_i} \in \mathbf{R}^{d_{\mathrm{model}}\times d_k}$、$\boldsymbol{W}^{K_i} \in \mathbf{R}^{d_{\mathrm{model}}\times d_k}$、$\boldsymbol{W}^{V_i} \in \mathbf{R}^{d_{\mathrm{model}}\times d_v}$ 为 3 组不同的线性矩阵。

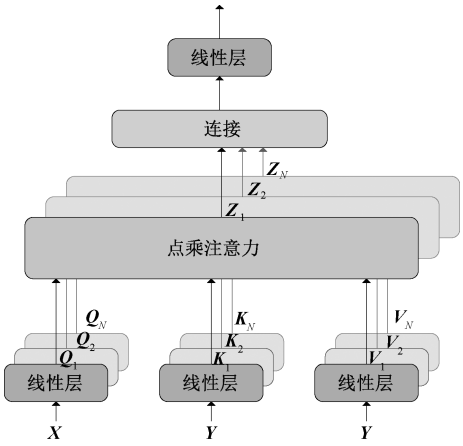


图 2 多头注意力结构图

Figure 2 Multihead attention structure

把此过程表述成一个统一的函数,如下所示:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。 \tag{5}$$

式中:自注意力权重由 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的点积运算生成;比例因子 $\sqrt{d_k}$ 和 Softmax 函数用于将自注意力权重归一化,所得到的权重被分配给 $\boldsymbol{V}$ 的相应元素,从而产生最终的输出向量。与卷积的稀疏连接类似,多头注意力将输入分离为具有 $\dfrac{d_{\mathrm{model}}}{h}$ 维向量的 h 个独立的注意力头,并行合并每个头的特征。在不增加额外计算量的情况下,丰富特征子空间的多样性。

1.4 前馈神经网络

前馈网络层(feed-forward network,FFN)等同于多层感知器(multilayer perceptron,MLP),主要由简单的神经网络组成,起到空间变化的作用。具体来说,对于多头注意力得到的特征再进行 FFN 非线性处理,可以挖掘特征的非线性关系,增强特征的表现能力。因此,多头注意力的输出被送入 2 个连续的带有 RELU 激活函数如下所示:

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{RELU}(W_1x + b_1)W_2 + b_2。 \tag{6}$$

式中:RELU(·)为激活函数;W 为权重;b 为偏置。FFN(·)由 2 个全连接层和激活函数组成,以此进行特征的维度变换。

1.5 图像分类结果

ViT 可在大规模数据集上做预训练,再迁移到小数据集上做微调,以此获得更好的图像分类结果,同时需要的计算资源也更少。ViT 使用由 3×10^8 张图片组成的大规模私有数据集 JFT-300 M 进行预训

练,在 ImageNet、CIFAR-10 和 CIFAR-100 图像分类数据集上进行测试,取得了与大多数主流 CNN 方法相似甚至更好的结果,实验结果如表 1 所示。由于 ViT 的注意力机制归纳偏置能力较弱,原始 patch embedding 为一个大卷积,难以获得底层信息,堆层数量受限,造成 ViT 的参数量和计算复杂度较大,且严重依赖于大规模数据集,因此,虽然已经证明了 ViT 在图像分类任务中的有效性,但 ViT 在训练数据不足的情况下很难推广,这也是下一步该方向的研究重点。

表 1　ViT 实验结果

Table 1　ViT test results

方法	数据集	准确度/%	参数量/M	复杂度/GFLOPs
ViT-B/16↑[5]	ImageNet	77.9	86	743
	CIFAR-10	98.1		
	CIFAR-100	87.1		
ViT-L/16↑[5]	ImageNet	76.5	307	5 172
	ImageNet	77.9		
	CIFAR-10	98.1		

2　改进 patch 模块

ViT 将图像分为多个不重叠的 patch,使 patch 间缺乏信息交互,而图像的边缘、线条和纹理等局部信息影响图像分类任务的准确性,从而降低实验性能。鉴于以上问题,Yuan 等[17] 提出 T2T-ViT(tokens-to-token Vision Transformers),该模型添加了局部性,每个 T2T 模块将相邻 token 聚合为一个 token,以此对周围 token 表示的局部结构进行建模,并可以缩短 token 长度,使得聚合相邻 token 后的大 token 具备局部性。通过实验发现,该模型对于 ViT 改进具备一定的优越性,但是由于重叠 patch 产生的冗余,该模型转换层的内存和计算负担沉重,还需要进一步进行实验探究。

Wu 等[18] 在 ViT 的 2 个主要部分引入卷积来提高模型性能和效率,主要为使用卷积投影操作替代原有的线性投影层以及在 ViT 前添加卷积 token 嵌入层,构造 CvT(convolutional Vision Transformers)。卷积 token 嵌入层对重构后的二维 token 特征图通过改变卷积运算参数调整每个阶段的 token 特征维度和数量。通过这种方式,在每个阶段逐步减少 token 序列长度,同时增加 token 特征维数,使得 token 能够在越来越大的空间上表示越来越复杂的视觉模式。而卷积投影层的目的是实现对局部空间上下文信息进行额外建模,并通过在多头注意力前使用深度可分离卷积对 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 矩阵进行下采样来提高效率。

Wang 等[19] 提出了具有金字塔结构的 ViT(pyramid Vision Transformer,PVT),其网络结构图如图 3 所示,该模型在 ViT 的基础上添加了金字塔结构,在 4 个阶段可以得到不同分辨率的特征图。并在多头注意力前采用空间缩减注意力层对 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 矩阵进行卷积来降低学习高分辨率特征图的资源成本,从而降低计算复杂度。PVT 在许多基线模型上证明了层次化 ViT 的可用性。Wang 等[20] 在 PVT 的基础上进行 3 种改进,改进后的模型称为 PVTv2。与 PVT 的非重叠 patch 分割方式不同,PVTv2 将 patch embedding 层修改为采用内核大小为 7×7、步长为 4 的卷积层,将图像分为可重叠的 patch,以加强 patch 间的联系;其次,在 FFN 层的 GELU 函数前添加深

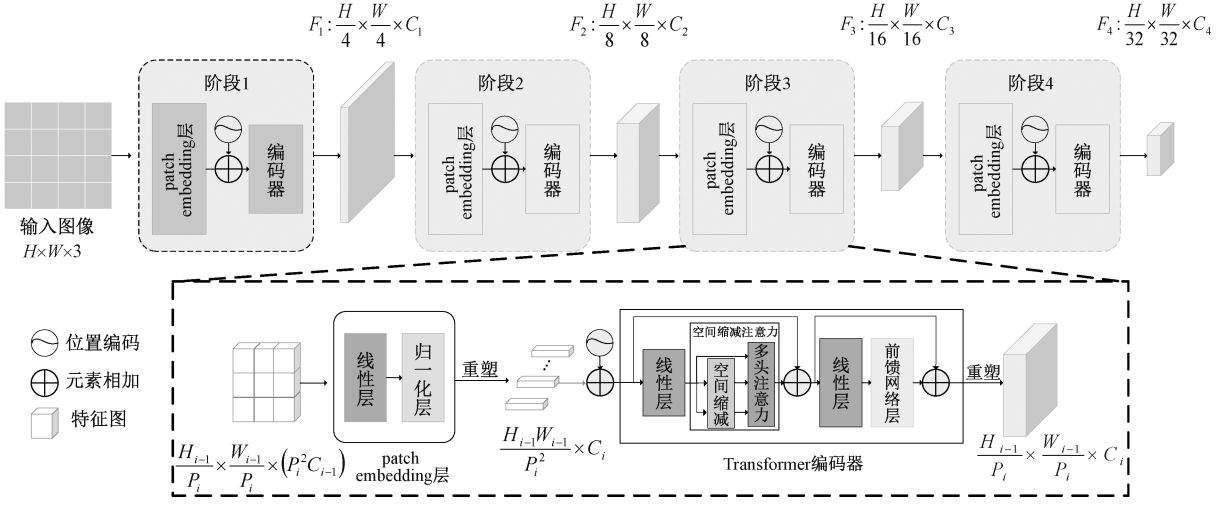


图 3　PVT 网络结构图

Figure 3　PVT network structure

度可分离卷积建模位置信息并且减少计算量,同时删除位置编码;最后将 PVT 的空间缩减注意力层中的卷积操作替换为对 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 矩阵进行平均池化操作,进一步降低模型的计算复杂度。通过以上修改,PVTv2 将计算复杂度降低到线性,并在图像分类任务中取得了很好的性能。

Pan 等[21]提出 LIT(less attention Vision Transformers),在早期阶段使用 MLP 对丰富的局部模式进行编码,并应用自注意力模块来捕获更深层中更长的依赖关系。同时,引入可变形卷积来改进 patch embedding 层。与常规卷积不同,可变形卷积可以对图像中包含信息量较多的区域进行特征提取,在此过程中调整特征提取位置的偏移量,以此提高模型的工作效率。

ViT 主要根据其结构或计算机视觉任务中基于多头注意力的模型进行重新设计。许多方法开始将 CNN 的层次结构或深度结构扩展到 ViT。上述模型共同的改进动机是将层次结构转移到 ViT,从而提高模型的泛化能力。所总结模型的实验结果在表 2 中给出。不难发现,通过改进 patch 模块不仅可以降低 ViT 的参数量和计算复杂度,还可以提高实验精度,使模型具有泛化性。

表 2 改进 patch 模块的方法对比

Table 2 Comparison of methods to improve the patch module

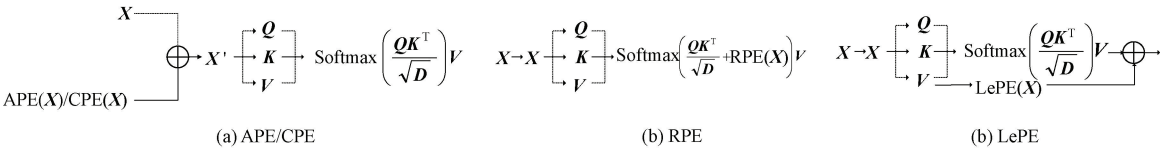
方法	轮次	批大小	参数量/M	复杂度/GFLOPs	准确度/%	优点	缺点
T2T-ViT-14[17]	310	512/1 024	21.5	4.8	81.5	强大的表示能力,并行计算,可解释性	对输入尺寸敏感,数据依赖性,难以处理长距离依赖关系
T2T-ViT-14↑384[17]			21.5	17.1	83.3		
T2T-ViT-24[17]			64.1	13.8	82.3		
CvT-13[18]	300	2 048	20.0	4.5	81.6	综合了卷积和 Transformer 的优势,处理多尺度信息,并行计算	参数量较大,对输入尺寸敏感,训练复杂度高
CvT-21[18]			32.0	7.1	82.5		
CvT-13↑[18]			20.0	16.3	83.0		
CvT-21↑[18]			32.0	24.9	83.3		
CvT-W24↑[18]			277.0	193.2	—		
PVT-Tiny[19]	300	128	13.2	1.9	75.1	处理多尺度信息,高性能,灵活性强	计算复杂度高,参数量较大,依赖于大规模数据集
PVT-Small[19]			24.5	3.8	79.8		
PVT-Medium[19]			44.1	6.7	81.2		
PVT-Large[19]			61.4	9.8	81.7		
PVTv2-B0[20]	300	128	3.4	0.6	70.5	更好的性能,高效的特征提取能力,多尺度特征融合	需要更多的计算资源,依赖于大规模数据集
PVTv2-B1[20]			13.1	2.1	78.7		
PVTv2-B2[20]			25.4	4.0	82.0		
PVTv2-B3[20]			45.2	6.9	83.2		
PVTv2-B4[20]			62.6	10.1	83.6		
PVTv2-B5[20]			82.0	11.8	83.8		
LIT-Ti[21]	300	1 024	19.0	3.6	81.1	减少计算复杂性,减少参数量	降低了模型的全局感知能力,对于复杂任务的性能影响不确定
LIT-S[21]			27.0	4.1	81.5		
LIT-M[21]			48.0	8.6	83.0		
LIT-B[21]			86.0	15.0	83.4		

3 改进位置编码

由于标准自注意力是置换变量,从而忽略 token 的位置信息,因此在 ViT 中用位置编码将此类位置信息添加回来。典型的位置编码机制包括绝对位置编码(absolute position encoding,APE)[22]、相对位置编码(relative position encoding,RPE)[4]和条件位置编码(conditional position encoding,CPE)[23]以及局部增强位置编码(locally enhanced position encoding,LePE)[24]。Shaw 等[22]使用的绝对位置编码是应用最广泛的一种编码。在 ViT 中,编码是用不同频率的正弦函数生成的,然后将它们添加到输入中。位置编码也是可学习的,使用固定维度矩阵实现,并使用系统自带的优化器与模型参数联合更新。Vaswani 等[4]提出了一种用于图像分类的二维相对位置编码,显示了相对于二维正弦嵌入的优越性。相对位置编码考虑了输入序列中 token 之间的距离。与绝对编码相比,相对位置编码具有平移不变性,并且可以在训练期间自然地处理比最长序列还长的序列。

在使用 Transformer 模型处理图像分类任务时,APE 和 RPE 处理能力受限,一方面限制了模型处理

比训练期间更长序列数据的能力,另一方面在每一个图像块上都添加位置编码会干扰模型的平移不变性。Chu 等[23]提出的 CPVT(conditional positional Vision Transformers)网络尝试使用 CPE 解决上述两个问题。与传统的位置编码方式不同,CPE 随输入大小而改变,可以为任意输入分辨率的特征图生成位置编码,在保持平移不变性的同时,也可以很容易地处理更长的输入序列,提高图像分类性能。而 CSWin Transformer[24]采用的 LePE 与 CPE 具有相同的思想,将位置编码作为并行模块添加到自注意力操作中,并对每个 ViT 块中的 $\boldsymbol{V}$ 进行深度卷积操作,以增强图像的局部信息。对比了以上 4 种位置编码机制如图 4 所示,由图 4 可以看出,APE 和 CPE 在输入 ViT 块之前将位置信息添加到 token 中,而 RPE 和 LePE 将位置信息合并到每个 ViT 块中,其中 LePE 以一种更直接的方式,将位置信息施加到 $\boldsymbol{V}$ 上,显示出更大的潜力,可以用作以后位置嵌入的改进方向。



(a) APE/CPE　　(b) RPE　　(b) LePE

图 4　位置编码对比图

Figure 4　Comparison diagram of position coding

将用于图像分类任务的无位置编码和 4 种位置编码机制进行总结比较如表 3 所示。根据消融实验结果可知,位置编码可以通过引入局部归纳偏差提高实验精度,虽然 RPE 在固定输入分辨率的分类任务中取得了和无 PE 相似的性能,但 LePE 在任意输入分辨率的任务上表现更好。因此,与 APE 和 CPE 相比,LePE 具备一定的优越性。

表 3　位置编码消融实验

Table 3　Position code ablation experiment

位置编码	准确度/%
无 PE	82.5
APE[22]	82.6
CPE[23]	82.2
RPE[21]	82.7
LePE[24]	82.7

4　改进多头注意力

4.1　基于窗口注意力的改进方法

ViT 在图像处理中面临视觉变化、高分辨率和大量像素的问题,导致其在不同场景下的适应性较差。同时,全局多头自注意力的使用增加了计算量。因此,Liu 等[25]提出了 Swin Transformer,利用沿空间维度的移位窗口机制来模拟全局和边界特征。该模型中金字塔式的层次结构用于缩减分辨率和扩大通道数,从而生成不同尺度的特征图,以便将其应用到计算机下游视觉任务中。Swin Transformer 把多头注意力限制在局部窗口中,再将 2 个连续的窗口进行移动,实现跨窗口交互,此过程将多头注意力层的计算复杂度从 $O(2n^2C)$ 降低到 $O(4M^2nC)$,其中 n 和 M 为 patch 大小和窗口大小。通过实验证明,Swin Transformer 在 ImageNet-1k 上达到了 84.2% 的分类精度。

与上述模型类似,Dong 等[24]提出了 CSWin Transformer,引入十字形窗口自注意力机制,用于并行计算水平和垂直条带形成的十字形窗口自注意力。每个条带通过将输入特征分割为等宽条带获得,并在网络的不同层改变条带宽度,在限制计算成本的同时实现强大的建模能力,同时引入 LePE 更好地处理局部位置信息。结合这些设计和分层结构,CSWin Transformer 在视觉任务上表现出和 Swin Transformer 相当的性能。

为了提高模型的效率,局部自注意力在局部区域内进行自注意力计算,造成在单个注意力层中的感受野不够大,导致上下文建模不足。当观察一个场景时,人类通常关注一个局部区域,同时以粗粒度关注非注意力区域。基于这一发现,Zhang 等[26]将 Swin Transformer 与 CSWin Transformer 的思想结合,设计了局部-全局交互的轴向扩展窗口自注意力机制,来解决 ViT 计算复杂度过高的问题,同时提高模型在高分辨率视觉任务的泛化性。轴向扩展窗口多头自注意力机制在局部窗口内执行细粒度多头自注意力,在水平轴和垂直轴上执行粗粒度多头自注意力,将 K 个头分成 3 个平行组,头数分别为 $\frac{K}{4}$、$\frac{K}{4}$ 和 $\frac{K}{2}$,将这 3 个并行组的输出重新连接,可以有效地捕获短期和长期视觉依赖关系,从而解决了 ViT 计算复杂度过高的问题,并大幅度提高图像分类精度。

由于多头自注意力机制在图像大小方面缺乏可扩展性,限制了其在最先进的视觉主干网络中的广

泛应用。Tu 等[27]提出了一个新的架构 MaxViT(multi-axis Vision Transformer),是一种高效、可扩展的多头自注意力模型,称之为多轴多头自注意力,包括局部窗口多头自注意力和网格多头自注意力两部分。窗口多头自注意力和 Swin Transformer 模型中的多头自注意力类似,都使用固定窗口大小来划分特征图,以获取局部信息。然而,网格自注意力则不同,将输入张量进行网格化,然后进行自注意力计算,以获取全局信息,从而平衡局部和全局计算复杂度。在自注意力块内部,窗口自注意力的输出作为网格自注意力的输入,使得每个块都可以实现局部和全局的空间交互,适应不同大小分辨率的输入。另外,通过结合自注意力模型和深度可分离卷积,可以减少参数数量,提高模型的泛化能力。同时,在多个阶段中重复使用基本构建块来组成分层主干网络。该模型在 ImageNet-1k 数据集上进行了实验,结果表明,MaxViT 与其他类似大小的典型模型相比,具备更少的参数量和更低的计算复杂度以及更高的实验精度,由此证明了该模型的优越性。

Fang 等[28]为缓解 ViT 效率和灵活性之间的冲突,为每个区域提出一个特定 token,作为“信使”(messenger,MSG)。通过操纵这些 MSG token,可以灵活地在不同区域交换视觉信息,降低计算复杂度。随后将 MSG token 整合到 MSG-Transformer 多尺度架构中。在图像分类任务中,MSG-Transformer 取得了具有竞争力的性能,并且在 GPU 和 CPU 上的推理速度都有所提升。

4.2 基于融合 CNN 和注意力的改进方法

虽然 ViT 在主流分类数据集上取得了与 CNN 相似甚至更好的性能,但在执行下游任务时,需要多尺度特征,而基本的 ViT 模型无法提供这种能力。此外,模型的计算量与输入图像大小呈二次复杂性,这使得计算开销非常大。因此,许多研究都致力于改进 ViT 模型以解决上述问题。

Han 等[29]利用 Transformer-in-Transformer(TNT)模型来聚合 token 和像素级表示。该模型使用 2 个 ViT,内部 ViT 模拟每个 patch 内的像素级交互,外部 ViT 提取全局信息。2 个 ViT 中间由一个线性投影层连接,该层将像素映射到其对应的 patch 来增强局部特征,以此得到局部与全局信息,消除 ViT 只关注全局信息所带来的弊端。

鉴于 Swin Transformer 模型中移位窗口的多头自注意力机制较为复杂,在现代深度学习中框架支持性较差,Chu 等[30]提出一个局部-全局分离 ViT 模型 Twins。Twins 用空间可分离的自注意力机制取代了复杂的 Swin Transformer 设计。在局部窗口内部使用 Swin Transformer 中的局部窗口自注意力进行计算,并对每个窗口内部的特征进行压缩,再使用类似于深度卷积或窗口式 TNT 块的全局自注意力机制去捕获各个窗口的关系。局部注意力层聚合每个子窗口内的相邻 patch 以增强细粒度特征,全局子采样注意力层用于捕获长距离特征。每个局部 patch 仅与其他 patch 及其相应的二维空间相邻交互。虽然该模型形式简单,但实现了与典型 Swin Transformer 的竞争。

Fan 等[31]设计了一种轻量级 ViT(CloFormer),利用上下文感知的局部增强模块,并采用双分支设计结构。所提出的卷积注意力有效地融合了共享权重和上下文感知权重,以聚合高频的局部信息。局部分支中卷积注意力首先使用具有共享权重的深度卷积提取局部表示。其次,使用上下文感知权重来增强局部特征;此外,卷积注意力将卷积算子应用于 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 以聚合局部信息,然后计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的哈达玛积,并对结果进行一系列变换,生成在[-1,1]的上下文感知权重。全局分支中则使用了传统的注意力机制,但对 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 进行了下采样以减少计算量,从而捕捉低频全局信息。CloFormer 能够同时发挥共享权重和上下文感知权重的优势,提高其局部感知的能力,使其在分类任务上取得了优异的性能。

如表 4 所示,大多数结构的改进方法都针对特定的模型大小、问题或特定的输入分辨率对模型进行了优化,从而降低模型的计算复杂度,使模型更易于训练,提高模型性能。

5 改进前馈神经网络

FFN 层尽管结构简单,但作为 ViT 模型的一个模块,在一定程度上影响其分类性能,传统的 FFN 结构示意图如图 5 所示。Shaw 等[22]提出简单地堆叠多头注意力模块会导致秩崩溃问题,造成 token 一致性电感偏差,而 FFN 层恰好能解决这一问题,成为 ViT 中不可缺少的一部分,因此出现了主要针对 FFN 模块的修改对 ViT 改进方法。

Guo 等[32]立足于 CNN 和 ViT 的交叉点,提出了一种新的用于视觉识别的模型 CMT(convolutional neural networks meet Vision Transformers)。输入图像首先经过卷积层进行细粒度特征提取,再送入多个堆叠的 CMT 块进行学习。CMT 是 ViT 的改进变体,其局部信息通过深度卷积得到增强,与 ViT 相比,CMT 第一阶段生成的特征图可以保持更高的分辨率,对于其他密集预测任务至关重要。此外,CMT

表 4　改进多头注意力机制方法对比

Table 4　Comparison of methods to improve multihead attention mechanism

方法	轮次	批大小	参数量/M	复杂度/GFLOPs	准确度/%	优点	缺点
Swin-T[25]	300/60	1 024/4 096	29.0	4.5	81.30	高效的注意力计算,高性能的视觉特征表示,良好的可扩展性	需要更多的计算资源,参数量较大
Swin-S[25]			50.0	8.7	83.00		
Swin-B[25]			88.0	15.4	83.30		
Swin-B ↑[25]			88.0	47.0	84.20		
Swin-L ↑[25]			197.0	103.9	—		
TNT-S[29]	300	1 024	23.8	5.2	81.30	强大的建模能力,优秀的空间感知性能,可解释性强	计算复杂度较高,参数量较大
TNT-B[29]			65.6	14.1	82.80		
TNT-S ↑[29]			23.8	—	83.10		
TNT-B ↑[29]			65.6	—	83.90		
Twins-S[30]	300	1 024	24.0	2.8	81.30	并行计算效率高,能够进行多尺度特征融合,可解释性强	训练难度较大,参数量较大
Twins-B[30]			56.0	8.3	83.10		
Twins-L[30]			99.2	14.8	83.30		
CSWin-T[24]	300	1 024	23.0	4.3	82.70	减少计算量,实验精度较高	感受野不够大,实验性能易受影响
CSWin-S[24]			35.0	6.9	83.60		
CSWin-B[24]			78.0	15.0	84.20		
CSWin-B[24]			173.0	31.5	85.40		
MaxViT-T[27]	300	4 096	31.0	33.7	85.72	可以实现局部与全局之间的空间交互,可适应不同大小分辨率的输入	参数量较大,需要更多的计算资源
MaxViT-S[27]			69.0	67.6	86.19		
MaxViT-B[27]			120.0	138.5	86.66		
MaxViT-L[27]			212.0	245.4	86.70		
AEWin-T[26]	300	256	23.0	4.0	83.60	能同时捕获局部与全局自注意力,参数量较小,计算复杂度降低	训练难度增大,参数量较多
AEWin-B[26]			78.0	14.6	85.00		
MSG-T[28]	300	1 024	25.0	3.8	82.40	灵活性高,速度快	参数量大,有时会丢失全局信息
MSG-S[28]			56.0	8.4	83.40		
MSG-B[28]			84.0	14.2	84.00		
CloFormer-XXS[31]	300	1 024	4.2	0.6	77.00	参数量少,计算复杂度低	实验精度需进一步提升
CloFormer-XS[31]			7.2	1.1	79.80		
CloFormer-S[31]			12.3	2.0	81.60		

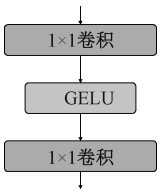


图 5　ViT 的 FFN 示意图

Figure 5　FFN diagram of ViT

采用类似于 CNN[33-35]的阶段式架构设计,使用步长为 2 的 4 个卷积层,逐层降低特征图分辨率并增加通道维数。通过此过程提取多尺度特征,并减轻高分辨率特征带来的计算负担。CMT 块中的局部感知单元(local perception unit,LPU)和反向残差前馈网络(inverted residual feed-forward network,IRFFN)可以帮助捕获中间特征中的局部和全局结构信息,提高网络的表示能力。其中 LPU 内部使用深度可分离卷积保持平移不变性,IRFFN 类似于反向残差块,由扩展层、激活函数层、深度卷积层和投影层以及归一化层组成,用于提取局部信息。通过上述改进提高其分类性能。

为了提高模型处理多尺度对象的能力,Ren 等[36]提出了一种新的策略,称为分流自注意力(shunted self-attention,SSA),允许 ViT 在每个注意力层的混合尺度上进行建模。SSA 的关键思想是将异质感受野大小注入 token,在计算注意力矩阵前,选择性地合并 token 以表示较大的对象特征,并保留某些 token 以保留细粒度特征。此合并方案可以使注意力学习不同对象大小之间的关系,减少 token 数量和计算成本。SSA 块与 ViT 中的传统自注意力块有 2 个主要区别,其一是 SSA 为每个注意力层引入了一种分流自注意力机制,通过对 $\boldsymbol{K}$、$\boldsymbol{V}$ 进行聚合得到不同尺度的特征图,以捕获多粒度信息和不同大小的更好的模型对象;其二是通过在 FFN 层中的激

活函数前增加一个深度可分离卷积层的残差连接增强跨 token 交互,提高局部建模能力。通过实验验证了该模型在 ImageNet 数据集上分类任务的有效性。

为了解决 ViT 对大规模数据集的依赖性,以及无卷积的 ViT 存在低层特征难提取和忽略空间维度局部性的问题,Yuan 等[37]提出了一种卷积增强图像 ViT(convolution-enhanced image Transformers, CeiT),结合了 CNN 提取低层特征以及 ViT 在处理长时依赖方面的优点。与 ViT 相比有 3 个方面的改进。第一,设计了一个图像到 token(image-to-tokens,I2T)模型,该模型从生成的低层特征中提取出较小尺寸的图像块,将图像块展平成一系列 token,而不是直接从原始输入图像中进行提取。受该结构的影响,I2T 模块没有引入更多的计算成本。第二,每个编码器块中的前馈网络层被替换为局部增强前馈(locally-enhanced feed-forward, LeFF)层,该层在空间维度上促进相邻 token 之间的相关性。第三,为了充分利用自注意力的能力,在 ViT 的顶部附加了一个利用多层表示的逐层 class token 自注意力(layer-wise class token attention, LCA)。该方法在 ImageNet 数据集上进行了实验,实验结果的提升有效验证了上述模型改进的有效性。

表 5 对本文针对 FFN 层的改进模型进行了分析比较。目前对 FFN 层改进较为常见的方法是在 FFN 层以串行或残差连接的方式添加深度可分离卷积,在降低模型参数量的同时增强模型对局部信息的建模能力,从而提高模型的分类精度。上述模型大多采用 300 轮、批大小为 1 024 的实验参数,在 ImageNet 数据集上进行相关实验,实验精度的提高有效佐证了对 FFN 层改进的可行性。

表 5 改进前馈神经网络方法对比

Table 5 Comparison of improved feed-forward neural network method

方法	轮次	批大小	参数量/M	复杂度/GFLOPs	准确度/%	优点	缺点
CMT-Ti[32]	300	1 024	9.5	0.6	79.1	CNN 捕获局部信息,Transformer 捕获全局信息	复杂度较高
CMT-XS[32]			15.2	1.5	81.8		
CMT-S[32]			25.1	4.0	83.5		
CMT-B[32]			45.7	9.3	84.5		
CMT-L[32]			74.7	19.5	84.8		
Shunted-T[36]	300	1 024	11.5	2.1	79.8	多尺度特征信息,高实验性能	参数量大,计算复杂度高
Shunted-S[36]			22.4	4.9	82.9		
Shunted-B[36]			39.6	8.9	84.0		
CeiT-T[37]	300	1 024	6.4	1.2	76.4	泛化能力较强,收敛速度较快,训练周期较短	引入了卷积层的参数,使得参数数量上升
CeiT-S[37]			24.2	4.5	82.0		
CeiT-T↑384[37]			6.4	3.6	78.8		
CeiT-S↑384[37]			24.2	12.9	83.3		

6 结束语

由于 ViT 依赖于大规模数据集,因此,如何对数据进行收集整理逐渐成为难点。同时 ViT 对模型进行训练时,过长的训练时间以及过高的计算复杂度使模型难以承担。鉴于此,越来越多的研究者关注到了小样本学习,以解决图像分类任务中数据集过大的问题,通过对 ViT 的不断研究,人们发现该模型中的部分组件并不是取得良好的图像分类结果的关键,鉴于 ViT 在图像分类任务的成功,逐渐将该方法迁移到视频分类任务。

(1)小样本学习。目前,基于深度神经网络的机器学习方法已经在人脸识别、自动驾驶、机器人等图像识别相关领域取得了巨大的成就,有的甚至已经超过人类目前的识别水平,但是还存在标注数据耗时和算力的问题。图像识别任务也因此受限,阻碍智能化图像识别技术的发展。而小样本学习是在少量标注数据上进行训练和学习,目前经常研究的问题为 N-way k-shot 形式,其含义为 N 类数据,每类数据包含 k 个标注样本。因此,如何将深度学习与少量样本的数据进行结合、基于小样本的图像识别研究逐渐成为关注的热点。

(2)ViT 未来改进方向。最近的一些方法[38-41]将 ViT 中的多头注意力机制抽象为一个 token 混合器,将 ViT 模型抽象为一个元模型,并通过向混合器里面填充不同的模块来探索注意力机制对 ViT 的影响。Yu 等[42]认为多头自注意力机制并不是影响模型性能的关键因素,因此,使用全局平均池化层替代自注意力机制取得了较高的实验精度,同时降低了计算量,使未来的研究方向不用过分关注于计算复杂度较高、参数量较大的自注意力机制。

(3)Video Transformer 发展方向。鉴于 ViT 在

图像分类的成功应用,该模型被逐渐应用于对视频数据进行建模。Video Transformer[43]与图像 ViT 的设计有共性,但视频固有的大维度将加剧 ViT 的局限性,需要进行特殊处理。同时,额外的时间维度还需要不同的嵌入、token 策略和相关架构。最后,视频媒体通常与其他模式配对,使得该模型特别容易在多模式设置中使用。

参考文献:

[1] 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.

[2] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. (2020-09-11)[2023-08-09]. https://arxiv.org/abs/1905.11946.

[3] 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.

[4] 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.

[5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2023-08-09]. https://arxiv.org/abs/2010.11929.

[6] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[J] Lecture Notes in Artificial Intelligence, 2020,12346: 213-229.

[7] WANG H Y, ZHU Y K, ADAM H, et al. MaX-DeepLab: end-to-end panoptic segmentation with mask Transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 5459-5470.

[8] CHENG B W, SCHWING A G, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation[EB/OL]. (2021-08-31)[2023-08-09]. https://arxiv.org/abs/2107.06278.

[9] CHEN X, YAN B, ZHU J W, et al. Transformer tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 8122-8131.

[10] JIANG Y F, CHANG S Y, WANG Z Y. TransGAN: two pure Transformers can make one strong GAN, and that can scale up[EB/OL]. (2021-12-09)[2023-08-09]. https://arxiv.org/abs/2102.07074.

[11] CHEN H T, WANG Y H, GUO T Y, et al. Pre-trained image processing Transformer[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 12294-12305.

[12] TAY Y, DEHGHANI M, BAHRI D, et al. Efficient Transformers: a survey[J]. ACM Computing Surveys, 2023, 55(6): 1-28.

[13] KHAN S, NASEER M, HAYAT M, et al. Transformers in vision: a survey[J]. ACM Computing Surveys, 2021, 54(S10): 1-41.

[14] HAN K, WANG Y H, CHEN H T, et al. A survey on Vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[15] LIN T Y, WANG Y X, LIU X Y, et al. A survey of Transformers[J]. AI Open, 2022, 3: 111-132.

[16] 毕莹, 薛冰, 张孟杰. GP 算法在图像分析上的应用综述[J]. 郑州大学学报(工学版), 2018, 39(6): 3-13.
BI Y, XUE B, ZHANG M J. A survey on genetic programming to image analysis[J]. Journal of Zhengzhou University (Engineering Science), 2018, 39(6): 3-13.

[17] YUAN L, CHEN Y P, WANG T, et al. Tokens-to-token ViT: training Vision Transformers from scratch on ImageNet[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 558-567.

[18] WU H P, XIAO B, CODELLA N, et al. CvT: introducing convolutions to Vision Transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 22-31.

[19] 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, 2021: 568-578.

[20] WANG W H, XIE E Z, LI X, et al. PVTv2: improved baselines with pyramid Vision Transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.

[21] PAN Z Z, ZHUANG B H, HE H Y, et al. Less is more: pay less attention in Vision Transformers[EB/OL]. (2021-12-23)[2023-08-09]. https://arxiv.org/abs/2105.14217.

[22] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[EB/OL]. (2018-04-12)[2023-08-09]. https://arxiv.org/abs/1803.02155.

[23] CHU X X, TIAN Z, ZHANG B, et al. Conditional positional encodings for Vision Transformers [EB/OL]. (2023-02-13) [2023-08-09]. https://arxiv.org/abs/2102.10882.

[24] DONG X Y, BAO J M, CHEN D D, et al. CSWin Transformer: a general Vision Transformer backbone with cross-shaped windows[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12114-12124.

[25] 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, 2021: 10012-10022.

[26] ZHANG Z M, GONG X. Axially expanded windows for local-global interaction in Vision Transformers[EB/OL]. (2022-11-13) [2023-08-09]. https://arxiv.org/abs/2209.08726.

[27] TU Z Z, TALEBI H, ZHANG H, et al. MaxViT: multi-axis Vision Transformer[C]//European Conference on Computer Vision. Cham: Springer, 2022: 459-479.

[28] FANG J M, XIE L X, WANG X G, et al. MSG-Transformer: exchanging local spatial information by manipulating messenger tokens[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12053-12062.

[29] HAN K, XIAO A, WU E H, et al. Transformer in Transformer[EB/OL]. (2021-08-26)[2023-08-09]. https://arxiv.org/abs/2103.00112.

[30] CHU X X, TIAN Z, WANG Y Q, et al. Twins: revisiting the design of spatial attention in Vision Transformers [EB/OL]. (2021-09-30)[2023-08-09]. https://arxiv.org/abs/2104.13840.

[31] FAN Q H, HUANG H B, GUAN J Y, et al. Rethinking local perception in lightweight Vision Transformer[EB/OL]. (2023-06-01)[2023-08-09]. https://arxiv.org/abs/2303.17803.

[32] GUO J Y, HAN K, WU H, et al. CMT: convolutional neural networks meet Vision Transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12165-12175.

[33] WOO S, DEBNATH S, HU R H, et al. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 16133-16142.

[34] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520.

[35] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11966-11976.

[36] REN S C, ZHOU D Q, HE S F, et al. Shunted self-attention via multi-scale token aggregation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10853-10862.

[37] 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.

[38] LEE-THORP J, AINSLIE J, ECKSTEIN I, et al. FNet: mixing tokens with Fourier Transforms[EB/OL]. (2022-05-26)[2023-08-09]. https://arxiv.org/abs/2105.03824.

[39] MARTINS A F T, FARINHAS A, TREVISO M, et al. Sparse and continuous attention mechanisms[EB/OL]. (2020-10-29)[2023-08-09]. https://arxiv.org/abs/2006.07214.

[40] MARTINS P H, MARINHO Z, MARTINS A F T. ∞-former: infinite memory Transformer[EB/OL]. (2022-05-25)[2023-08-09]. https://arxiv.org/abs/2109.00301.

[41] RAO Y M, ZHAO W L, ZHU Z, et al. Global filter networks for image classification[EB/OL]. (2021-10-26)[2023-08-09]. https://arxiv.org/abs/2107.00645.

[42] YU W H, LUO M, ZHOU P, et al. MetaFormer is actually what you need for vision[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10819-10829.

[43] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding?[EB/OL]. (2021-02-24)[2023-08-09]. https://arxiv.org/abs/2102.05095.

A Review of Vision Transformer for Image Classification

ZHI Min, LU Jingfang

(School of Computer Science and Technology, Inner Mongolia Normal University, Hohhot 010022, China)

Abstract: ViT as a model based on the Transformer architecture has shown good results in image classification tasks. In this study, the application of ViT on image classification tasks was systematically summarized. Firstly, the functional characteristics of the ViT framework and its four modules (patch module, position encoding, multihead attention mechanism and feed-forward neural network) were briefly introduced. Secondly, the application of ViT in image classification tasks was summarized with the improvement measures of the four modules. Due to the fact that different model structures and improvement measures could have a significant impact on the final classification performance, a side-by-side comparison of various types of ViTs was made in this paper. Finally, the advantages and limitations of ViT in image classification were pointed out, and possible future research directions were proposed to break the limitations, and further to extend the application of ViT in other computer vision tasks. The extension of ViT to a wider range of computer vision fields, such as video understanding, was explored.

Keywords: ViT model; image classification; multihead attention; feed-forward network layer; position encoding

---

(上接第 10 页)

Research of Mobile Edge Computing for Future Mobile Communications: A Review

YANG Shouyi[1], CHEN Yihang[1], ZHANG Shuangling[2], HAN Haojin[1], LI Guangyuan[3], HAO Wanming[1]

(1. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Department of Mechanical and Electrical Engineering, Henan Light Industry Vocational College, Zhengzhou 450002, China; 3. Department of Engineering, Huanghe University of Science and Technology, Zhengzhou 450061, China)

Abstract: Mobile edge computing (MEC) has become one of the key technologies for future-oriented communications by offloading the computing and storage tasks of mobile terminals from centralized data centers to edge grids to satisfy the diverse device service demands in complex communication scenarios. The basic concept and basic framework of MEC technology were introduced by describing the development history from cloud computing, fog computing to mobile edge computing. On this basis, the research progress of MEC was discussed in four aspects, namely, computation offloading, resource allocation, cache management, and security protection. A detailed overview of the relevant research results was provided. Then, studies on several typical application scenarios of edge computing such as IoT, MEC combined with blockchain, AI-assisted MEC systems, integrated sensing and communication, and cloud-edge collaboration were summarized. It demonstrated the potential benefits of MEC in 6G in terms of constituting an intelligent, efficient and secure communication network. Finally, the challenges faced by MEC research in convergence innovation from the aspects of interoperability, security risk, mobility management and scalability were pointed out, as well as the advantages and development trends in the directions of ultra-reliable low-latency communications, communication-sensing-computing integration and satellite-ground fusion mobile communication. The development trend of it in the future mobile communication was also summarized and outlooked.

Keywords: mobile communication; mobile edge computing; computation offloading; resource allocation; information security