

DF-Former: 一种空域与频域融合学习的图像分类方法

余坚毅, 蔡之华

(中国地质大学(武汉)计算机学院, 湖北 武汉 430072)

摘要: Transformer模型具备强大的低频全局信息提取能力,但是缺乏对高频局部信息的感知且受限于自注意力的高计算开销,从而严重限制了其实际应用。针对以上问题,提出以“分治—引导—融合”为核心的一种无需人类专家先验知识的高效频率分支结构SFF,并通过“先降维再升维”的方法,有效减少了计算量。SFF由空域编码器、频域编码器和域特征融合模块构成,能够高效提取并融合高频局部特征与低频全局特征。基于SFF框架,进而提出一种全新视觉网络DF-Former,在公开数据集ImageNet10和CINIC10的准确率分别为85.18%和91.32%,在参数数量和计算量低于传统模型的情况下,取得了良好的性能。

关键词: 图像分类; 频率引导; 双层编码; 视觉Transformer

DOI: 10.11907/rjtk.241672

开放科学(资源服务)标识码(OSID):

中图分类号: TP301

文献标识码: A

文章编号: 1672-7800(2025)009-0168-06



DF-Former: An Image Classification Method with Fusion Learning in Spatial and Frequency Domains

YU Jianyi, CAI Zhihua

(School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430072, China)

Abstract: The Transformer model possesses a robust capability for extracting low-frequency information but lacks sensitivity to high-frequency information and is constrained by the high computational cost of self-attention, which severely limits its practical applications. To address these issues, this paper proposes an efficient frequency-branched structure named SFF, which is centered around the concept of 'divide, guide, and fuse' and does not require prior knowledge from human experts. The SFF adopts the philosophy of 'dimensionality reduction followed by dimensionality expansion' to further reduce its computational load. Comprising a spatial encoder, a frequency encoder, and a domain feature fusion module, SFF is capable of efficiently extracting and fusing high-frequency local features and low-frequency global regional features. Based on the core of SFF, this paper further proposes a novel vision network called DF-Former. On the public datasets ImageNet10 and CINIC10, DF-Former achieves an accuracy of 85.18% and 91.32%, respectively, demonstrating excellent performance with a lower parameter count and computational load compared to traditional models.

Key Words: image classification; frequency guidance; dual encoding; visual Transformer

0 引言

图像分类算法作为计算机视觉四大核心任务之一,具有广泛的应用前景。自2012年AlexNet提出以来,卷积神

经网络凭借其优异的局部特征提取能力和固有的归纳偏置特性,逐渐成为计算机视觉领域的主流研究方法^[1]。然而,受限于滑动窗口的工作机制,传统卷积神经网络难以建立长距离依赖关系,这种对全局信息感知能力的缺失导致其分类精度面临瓶颈。

收稿日期: 2024-08-08

扫描二维码阅读全文:



基金项目: 国家自然科学基金项目(61973285); 湖北省自然科学基金项目(2022CFB)

作者简介: 余坚毅(2000-),男,CCF会员,中国地质大学(武汉)计算机学院硕士研究生,研究方向为深度学习与图像分类; 蔡之华(1964-),男,博士,中国地质大学(武汉)计算机学院教授、博士生导师,研究方向为人工智能、机器学习和演化算法等。本文通讯作者: 蔡之华。

2020年,ViT模型将自然语言处理领域中的Transformer模型引入计算机视觉领域^[2]。ViT通过自注意力机制具备了强大的长距离依赖关系捕捉能力,能够学习空间位置之间的相互作用,从而实现全局感受野。但由于缺乏归纳偏置,ViT训练收敛速度慢且信息利用不充分,其仅在大量数据训练时具备优势。此外,自注意力机制的时间复杂度高,其在面对高分辨率数据时所需计算资源高昂。

基于卷积神经网络和ViT优势互补的特点,混合架构模型将二者的优点进行了结合,通过将卷积引入ViT架构中,从而使模型具备了归纳偏置的能力,改善了分类正确率。基于此,本文提出了一种全新的混合架构分类模型DF-Former(Domain Fusion Former),实现了对全局和局部信息的融合感知。

1 相关工作

目前,主流深度学习图像分类算法可以分为纯卷积神经网络模型、纯Transformer架构模型和混合架构模型三大类。

纯卷积神经网络模型的代表模型为ResNet,其提出残差连接操作,成功地解决了深度神经网络中的退化问题^[3]。

纯Transformer架构模型的代表模型为Swin Transformer,其将自注意力改进为移动窗口注意力^[4]。移动窗口注意力机制通过将特征图划分为多个局部非重叠的小窗口,将自注意力的计算限制在这些小窗口内,从而实现了线性计算复杂度。

混合架构模型的代表模型是iFormer^[5]、ConvNeXt^[6]和InceptionNeXt^[7]。Former整体架构引入Transformer架构,特征提取部分同时采用了深度可分离卷积和自注意力机制,将通道手动划分为高频分支和低频分支。其中,大部分通道划分给低频分支,高频分支使用最大池化聚焦局部细节,低频分支使用平均池化整合特征图的全局模式。ConvNeXt是基于ResNet进行改进,整体架构引入Transformer架构,特征提取部分仅采用深度可分离卷积。InceptionNeXt则是使用Inception改进ConvNeXt。

由上述可知,Transformer模型的有效性核心在于其整体架构,而非仅仅依赖于自注意力机制。因此,分类模型的构建应当在整体架构上采用Transformer架构,而在特征提取方法的构建上则不必局限于自注意力机制。

神经网络倾向于找到简单的分类解决方案,其在训练中优先进行学习知识取决于当前最独特的频率特征,这些特征既可能是低频信息,也可能是高频信息。意味着神经网络在提取低频特征时,高频特征可能成为阻碍其学习效率的噪声,反之亦然^[8-9]。因此,局部特征和全局特征在分类模型的构建中同样重要,应当为其分配均衡的计算量。而上述代表工作中,Swin Transformer将自注意力改造

成为局部自注意力,虽然降低了时间复杂度,但是缺乏全局感受野;ConvNeXt和InceptionNeXt仅使用了卷积进行特征提取,缺乏对全局信息的感知;iFormer同时采用了卷积和自注意力,具备对局部和全局信息的感知,但是过多的计算量分配给了全局特征的提取。

ConvNeXt、InceptionNeXt和iFormer均使用了深度可分离卷积(DwConv)来降低计算量,虽然降低了理论FLOPS,但深度可分离卷积也带来了层数增加,伴随着运算符对内存的频繁访问,实际训练中并没有实现加速的效果。对此,FasterNet提出部分卷积Partial Convolution(PCConv),其能够在减少不必要计算和内存访问的情况下更有效地进行空间特征提取^[10]。因此,局部特征提取方法的构建可在部分卷积的基础上进行。

受到以上工作的启发,本文按照“分治—引导—融合”的思想提出一种高效的空域和频域特征融合学习的图像分类方法DF-Former。其中,“分治”将特征学习任务拆分为局部特征提取和区域全局特征提取2个子任务,以便进行精细化处理;“引导”按照不同子任务的要求来突出特征图的关键信息,以提高特征学习的效率;“融合”建立不同尺度特征的联系,以应对复杂的视觉场景变化,从而提升模型泛化性。DF-Former的核心是SFF(Spatial-Frequency Fusion)。SFF将通道均等地划分为高频分支和低频分支,并送入空域编码器SE和频域编码器FE,SFF分别引导神经网络在空域提取局部特征,在频域提取区域和全局特征,最后由域特征融合模块DF进行跨域信息的融合。空域编码器SE由最大池化、PCConv和空间移位操作组成;SFF由平均池化操作、频域层和上采样操作组成;域特征融合模块DF由逐点卷积和ECA注意力组成。

与ConvNeXt、InceptionNeXt和Swin Transformer相比,DF-Former具备对局部和全局信息的融合感知,对图像信息的利用更加充分。同时,DF-Former弃用了ConvNeXt和InceptionNeXt中的深度可分离卷积,从高效的部分卷积出发进行局部细节建模,有助于提升模型效率。和iFormer相比,DF-Former对局部特征提取和全局特征提取分配了均等的通道,从而避免了局部细节信息的丢失。在全局特征提取部分,iFormer采用了自注意力,而DF-Former则引入了频域层,这种方法在降低计算量的同时,实现了对全局关系的有效感知。

2 DF-Former模型

2.1 DF-Former整体架构

为了便于定量对比和后续研究,本文沿用了iFormer的宏观框架,仅将其核心算法模块iFormer Block中的Inception Mixer替换为本文提出的Spatial-Frequency Fusion,其余部分保持不变。DF-Former的整体架构如图1所示。

在DF-Former的总体流程中,特征图经过4个Stage的

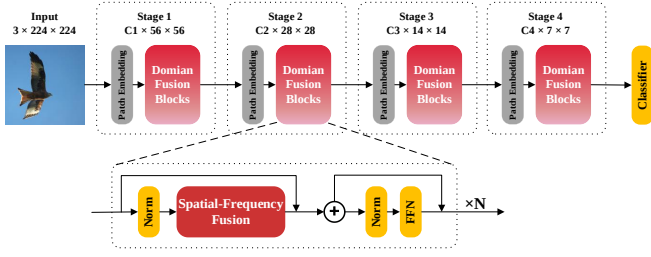


Fig. 1 DF-Former overall architecture

图1 DF-Former 整体架构

处理,特征图整体呈现金字塔架构,每个 Stage 串行多个 Domain Fusion Block 从空域和频域的角度联合对特征图进行特征的提取,并对域特征进行融合。由于图像的高频部分主要包含局部细节信息,而低频部分则更多地承载图像的整体轮廓和纹理模式,同时神经网络倾向于优先学习当前特征图中最独特频率的特征,这意味着在神经网络在学习低频特征时,高频特征可能会成为学习过程中的噪声,反之亦然。因此,DF-Former 将通道均等划分为高频分支和低频分支,使神经网络可以更加轻松地进行多尺度的特征提取,从而提升模型的计算效率。

2.2 SFF 结构

本文提出了空域与频域融合结构 (Spatial-Frequency Fusion, SFF),如图 2 所示。SFF 结构由空域编码器 (Spatial Encoder, SE)、频域编码器 (Frequency Encoder, FE)、域特征融合模块 (Domain Fusion, DF) 组成。以上子模块的输入和输出格式均为 $B \times C \times H \times W$ 格式。

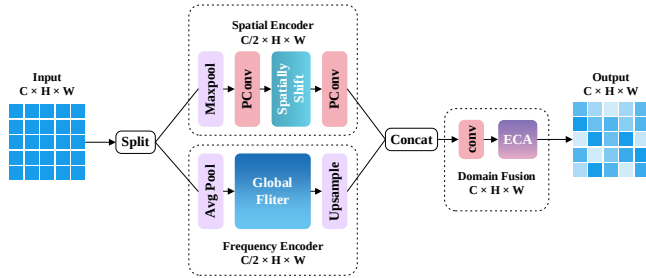


Fig. 2 Spatial and frequency domain fusion structure SFF

图2 空域和频域融合结构 SFF

SFF 的具体技术细节如下:首先,对于任意输入的一个特征图 $X \in \mathbb{R}^{C \times H \times W}$,SFF 按照其通道的维度均等地连续分割成 2 个分支。其中,一个分支是 $X_L \in \mathbb{R}^{C \times H \times W}$,另一个分支则是 $X_H \in \mathbb{R}^{C \times H \times W}$, X_L 和 X_H 分别各占前二分之一通道以及后续通道所对应特征图。其次,将 X_L 送入频域编码器 FE,将 X_H 送入空域编码器 SE 进行多尺度特征提取,频域编码器引导神经网络在频域捕捉低频的区域和全局特征 *Regional Global*,空域编码器引导神经网络在空域捕捉高频的局部特征 *Local*。再次,频域编码器和空域编码器完成特征提取后按照通道维度将 $Local \in \mathbb{R}^{C \times H \times W}$ 和 $RegionalGlobal \in \mathbb{R}^{C/2 \times H \times W}$ 连接 (Concat 操作) 成为一个连续的特征图 $(SE(X_H) + FE(X_L)) \in \mathbb{R}^{C \times H \times W}$,再将其送入域特征融合模块 DF,对空域和频域信息进行特征融合。这

些空域局部特征和频域区域全局特征互不干扰,相互融合后的特征使得后续阶段的空域编码器 SE 和频域编码器 FE 能够更有效地提取特征,最后,得到输出特征图 $Output \in \mathbb{R}^{C \times H \times W}$ 。SFF 的整体框架可以被归纳为如公式 (1)—公式 (4) 所示。

$$X \in \mathbb{R}^{C \times H \times W} \quad (1)$$

$$X_L = X[:, :, C/2, :, :] \quad (2)$$

$$X_H = X[:, :, C/2, :, :] \quad (3)$$

$$Output = DF(Concat(SE(X_H), FE(X_L))) \quad (4)$$

2.3 频域编码器 FE

针对传统自注意力存在的计算复杂度高、参数量大的问题,受 GFNet^[11] 的频域捕获特征思想的启发,本文弃用自注意力机制,设计了一种高效区域全局特征学习模块频域编码器 FE,其在频率域引导神经网络对区域和全局特征进行捕获。

具体而言,本文使用二维离散傅里叶变换将图像从空域映射到频域。对于任意给定的一个长度为 N 的复数序列 $S[n]$,通过一维离散傅里叶变换 (1D DFT),可以将其转换到频域,具体如公式 (5) 所示。

$$F[k] = \sum_{n=0}^{N-1} S[n] e^{-i(2\pi/N)kn} \quad (5)$$

其中, i 是虚数单位, $F[k]$ 表示当序列频率为 $\varphi = (2\pi/N)kn$ 时的频谱。相应地, DFT 的逆操作 IDFT 如公式 (6) 所示。

$$S[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] e^{i(2\pi/N)kn} \quad (6)$$

上述 1D DFT 可以自然地推广到 2D DFT。对于任意形状为 (M, N) 的二维实数信号 $S[m, n]$,可以通过公式 (7) 将其转换到频域。

$$F[u, v] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} S[m, n] e^{-i\left(\frac{um}{M} + \frac{vn}{N}\right)2\pi} \quad (7)$$

本文使用快速傅里叶变换 (FFT) 算法来计算 DFT 和 IDFT。相较于自注意力 $O(N^2)$ 规模的计算复杂度,FFT 利用频率信息的对称性和周期性将计算复杂度降低到了 $O(N \log N)$ 。此外,在本文中的 FFT 归一化方式选择的是正交归一化,其优势如下:①变换后的频谱在频率域中分布均匀,且能够保证能量守恒(转换时信号的能量不会丢失或增加);②正交变换所产生的是不包含冗余信息压缩的频谱表示,这意味着频域中的每个频率分量都对应原始信号中的一个独特频率成分,频谱之间不会相互重叠,从而有利于频率成分的分离,并避免信号重建时的失真。

本文采用“先降维再升维”的策略,从而进一步降低 DFT 的计算复杂度,具体操作如下:①对输入特征图进行平均池化,该操作对空间尺寸进行了下采样,实现了信息的抽象概括,使得神经网络能够更高效的捕捉区域和全局信息;②通过 DFT 将降维后的特征图从空域映射到频域,将其与可学习参数按元素相乘,得到频域学习特征;③通

过DFT的逆操作将频域学习特征恢复到空域,再通过上采样操作恢复空间尺寸便得到了低频特征。

频域编码器FE的技术实现细节如下:首先,对于任意给定输入的一个特征图 $L \in \mathbb{R}^{C \times H \times W}$,经过平均池化操作(其中窗口大小 pool_size 设置为2, stride 步长设置为2,填充 padding 设置为0,且计算平均值时包含填充的像素)得到一个聚焦区域和全局的特征 $L_A \in \mathbb{R}^{C \times H/2 \times W/2}$;其次,进行二维离散傅里叶变换得到频域特征 $L_F \in \mathbb{C}^{C \times H/2 \times (W/4 + 1)}$;再次,设置一个和 L_F 同样形状的复数类型可学习参数 K (不包括批量 Batch 这个维度),将 L_F 和 K 按元素相乘得到 $L_K = L_F \odot K \in \mathbb{C}^{C \times H/2 \times (W/4 + 1)}$;最后,进行逆二维离散傅里叶变换恢复到空域,并进行 Upsample 上采样操作(其中 scale factor 设置为 pool size 大小)得到 $RegionalGlobal \in \mathbb{R}^{C \times H \times W}$ 。

FE算法的框架可归纳为公式(8)—公式(13)。

$$L \in \mathbb{R}^{C \times H \times W} \quad (8)$$

$$L_A = \text{Average Pooling}(L) \in \mathbb{R}^{C \times H/2 \times W/2} \quad (9)$$

$$L_F = \text{FFT}(L) \in \mathbb{R}^{C \times H/2 \times W/2} \quad (10)$$

$$L_K = L_F \odot K \in \mathbb{C}^{C \times H/2 \times (W/4 + 1)} \quad (11)$$

$$RG = \text{Upsample}(\text{IRFFT2D}(L_K)) \quad (12)$$

$$\text{Regional Global} = RG \in \mathbb{R}^{C \times H \times W} \quad (13)$$

2.4 空域编码器SE

卷积神经网络具备强大的局部上下文提取能力,受FasterNet和ShiftViT空间移位思想的启发,本文设计了一种高频分支模块空域编码器(Spatial Encoder, SE),能够高效地从空间域引导神经网络拟合高频局部细节。

空域编码器的技术实现细节如下:采用Partial Conv ($PConv$)为模型引入归纳偏置。 $PConv$ 通过有效利用特征图在通道间的高度相似性,实现了高效的局部特征提取。具体而言, $PConv$ 选取第一个或最后一个连续的四分之一通道作为整个通道特征的代表,对这部分输入通道使用一个卷积核大小为3×3、stride为1、padding为1的普通卷积进行空间局部特征提取。此外,该卷积操作不使用偏置,以减少模型参数、防止过拟合并提升泛化能力。相较于深度可分离卷积(DwConv), $PConv$ 可以在大幅减少计算量和推理时间的情况下,高效地进行局部特征的提取。

基于 $PConv$,本文提出了更高效的局部特征提取模块—空域编码器SE。首先,对于任意给定的输入特征图 $H \in \mathbb{R}^{C \times H \times W}$,进行Max Pooling最大池化操作(窗口大小 kernel_size 设置为3,步长和填充均设置为1),得到局部细节聚焦后的特征图 $H_{MP} \in \mathbb{R}^{C \times H \times W}$,此时局部高频成为神经网络中最独特的频率,神经网络将优先提取局部细节;其次,将聚焦后的特征图送入 $PConv$ 进行局部特征提取得到 $H_{pc} \in \mathbb{R}^{C \times H \times W}$;再次,对 H_{pc} 进行Spatially Shift空间移位操作^[12],进一步融合局部空域信息得到 $H_{ss} \in \mathbb{R}^{C \times H \times W}$;最后,将 H_{ss} 送入 $PConv$ 再进行一次局部特征提取得到 $Local \in \mathbb{R}^{C \times H \times W}$ 。

其中,Spatially Shift空间移位操作的流程如下:首先,

沿着通道维度选择最先或最后一个连续的三分之一通道,分割得到待移位的特征图,并对这部分特征图进行空间移位变换,shift的步长设置为1个像素,其余特征图保持不变;其次,将这部分特征通道均分为4份特征图,分别沿左移1个像素、右移1个像素、上移1个像素和下移1个像素;最后,在作用域外的像素被简单删除,空的像素被零值填充。SE算法的框架可归纳为公式(14)—公式(18)。

$$H \in \mathbb{R}^{C \times H \times W} \quad (14)$$

$$H_{MP} = \text{Max Pooling}(H) \in \mathbb{R}^{C \times H \times W} \quad (15)$$

$$H_{PC} = PConv(H_{MP}) \in \mathbb{R}^{C \times H \times W} \quad (16)$$

$$H_{SS} = \text{Spatially Shift}(H_{PC}) \in \mathbb{R}^{C \times H \times W} \quad (17)$$

$$Local = PConv(H_{SS}) \in \mathbb{R}^{C \times H \times W} \quad (18)$$

2.5 域特征融合模块DF

在空域编码器SE和频域编码器分支FE中,神经网络分别被引导对高频特征和低频特征进行高效的多尺度提取。通过特征相加Concat操作,本文提出了域特征融合模块(Domain Fusion, DF),用于融合SE和FE处理后的特征。该模块实现了空域信息和频域信息的跨域特征融合,从而使后续的SFF模块能够进行更精准的决策。

域特征融合模块DF的具体技术细节如下:首先,对任意给定的输入特征图 $SpatialFrequency \in \mathbb{R}^{C \times H \times W}$,通过1×1卷积(步长为1,填充为0)进行空间信息融合;其次,送入ECA注意力(Efficient Channel Attention)^[13]通道特征进行加权融合;最后,得到特征表示 $Fusion \in \mathbb{R}^{C \times H \times W}$ 。DF算法的框架可归纳为公式(19)—公式(21)。

$$SF = \text{SpatialFrequency} \in \mathbb{R}^{C \times H \times W} \quad (19)$$

$$Fusion = \text{ECA}(PWConv(SF)) \quad (20)$$

$$Fusion \in \mathbb{R}^{C \times H \times W} \quad (21)$$

3 实验方法与结果分析

3.1 基准数据集

本文选取了ImageNet-100和CINIC-10作为基准测试数据集,以Top-1和Top-5准确率作为精确度评价指标。其中,Imagenet100是ImageNet-1K的子集,包含100个类别、10万张训练图像、2万张验证图像以及5000张测试图像^[14];CINIC-10是ImageNet和CIFAR-10的融合数据集,其包含10个类别、9万张训练图像、9万张验证图像以及9万张测试图像^[15]。

3.2 实验环境和参数设置

本文所有实验均迭代了310个epoch,在单张NVIDIA Tesla V100 32G上进行训练。将数据集图像的像素分辨率统一调整为224×224作为模型单个数据的输入尺寸,批量数据处理大小Batch_size详见表1。使用平滑技术smoothing,使用channels last内存格式进行无损加速训练,学习率设置为5e-4,最小学习率设置为5e-6,学习率预热周期Warmup epochs设置为30,预热warmup lr设置为1e-

6,开始衰减学习率的周期数 decay epochs 设置为 30, 衰减率 weight decay 设置为 $5e-2$, 使用余弦学习率调度器。数据预处理时使用的均值 mean 设置为 0, 标准差 std 设置为 0.02, 动量 momentum 设置为 0.9, DropPath 比例设置为 0.2。数据增强概率的参数设置如下: cutmix 的概率设置为 0.4, mixup 的概率设置为 0.5, 水平翻转 hflip 的概率设置为 0.5, 垂直翻转 vflip 的概率设置为 0.5, 颜色抖动 color jitter 的概率设置为 0.5。

Table 1 Results of comparison experiment

表 1 对比实验结果

Dataset	Model	Batch_ size	Params /M	FLOPS /G	Top-1 /%	Top-5 /%
ImageNet100	ResNet50	296	23.71	4.11	83.74	95.58
	Swin-Tiny	252	27.59	4.51	83.82	95.44
	InceptionNeXt-T	168	25.77	4.20	85.37	96.24
	IFormer-S	168	19.52	4.85	85.72	96.46
	DF-Former	214	16.23	3.95	85.18	96.44
	ResNet50	296	23.71	4.11	80.58	98.25
CINIC-10	Swin-Tiny	252	27.59	4.51	79.12	98.31
	InceptionNeXt-T	168	25.77	4.20	87.39	99.23
	IFormer-S	168	19.52	4.85	83.61	98.07
	DF-Former	214	16.23	3.95	91.32	99.78

3.3 对比实验

对比实验结果(见表 1)表明,在数据集 ImageNet100 上,DF-Former 模型展现出显著的性能优势。与 ResNet50 相比,DF-Former 在参数量减少 31.55%、计算量(FLOPS)降低 3.9% 的同时,Top-1 准确率提升了 1.44%,Top-5 准确率提升了 0.86%;与 Swin-Tiny 相比,DF-Former 在参数量减少 41.17%、FLOPS 降低 12.42% 的情况下,Top-1 准确率提升了 1.36%,Top-5 准确率提升了 1%;与 IFormer-S 相比,DF-Former 在参数量减少 16.85%、FLOPS 降低 18.56% 的条件下,Top-1 准确率仅下降 0.54%,Top-5 准确率仅下降 0.02%;与 InceptionNeXt-T 相比,DF-Former 在参数量减少 37%、FLOPS 降低 6% 的情况下,Top-1 准确率仅下降 0.19%,Top-5 准确率提升了 0.2%。

在低分辨率数据集 CINIC-10 上,DF-Former 则是超越了全部基准模型。与 ResNet50 相比,DF-Former 在参数量减少 31.55%、FLOPS 降低 3.9% 的同时,Top-1 准确率提升了 10.74%,Top-5 准确率提升了 1.53%;与 Swin-Tiny 相比,DF-Former 在参数量减少 41.17%、FLOPS 降低 12.42% 的情况下,Top-1 准确率提升了 12.2%,Top-5 准确率提升了 1.47%;与 InceptionNeXt-T 相比,DF-Former 在参数量减少 37%、FLOPS 降低 6% 的条件下,Top-1 准确率提升了 3.93%,Top-5 准确率提升了 0.55%;与 IFormer-S 相比,DF-Former 在参数量减少 16.85%、FLOPS 降低 18.56% 的情况下,Top-1 准确率提升了 7.71%,Top-5 准确率提升了 0.71%。

与 ResNet50、Swin-Tiny 和 InceptionNeXt-T 相比,DF-Former 不仅能够捕捉局部细节,还能够对全局关系进行建模,从而实现了更高的分类精度。此外,与 InceptionNeXt-

T 和 IFormer-S 相比,DF-Former 弃用了深度可分离卷积,在部分卷积基础上进行局部关系建模,其参数量更低、计算效率更高。与 IFormer-S 相比,DF-Former 弃用了自主注意力,转而采用频域层来建模全局关系,进一步提升了计算效率。

3.4 消融实验和可解释性分析

为了探究 DF-Former 内在结构对模型性能的影响,本文结合消融实验和 GradCAM 方法对其进行可解释性分析。表 2 为消融实验结果,其中,SE-only 表示去掉了 FE 模块,仅使用空域编码器 SE 进行特征提取;FE-only 表示去掉了 SE 模块,仅使用频域编码器 FE 进行特征提取;No-ECA 表示去掉了 ECA 通道注意力。消融实验结果验证了各个子模块的有效性:与 DF-Former 相比,仅采用局部特征学习的 SE-only 模型准确率下降了 0.28%;No-ECA 模型准确率下降了 0.58%;而仅采用区域全局特征学习的 FE-only 模型准确率下降了 3.7%。

Table 2 Ablation experiment results

表 2 消融实验结果

Dataset	SE	FE	ECA	Params/M	FLOPS/G	Top-1/%	Models
ImageNet-100	×	√	√	16.02	3.85	81.48	FE-only
	√	×	√	16.87	4.27	84.90	SE-only
	√	√	×	16.23	3.95	84.60	No-ECA
	√	√	√	16.23	3.95	85.18	DF-Former

图 3(a)是真实环境下的蜗牛图像样本,蜗牛整体形状和颜色具备环境隐蔽的特点,这要求模型具备良好的鲁棒性。图 3(b)、图 3(c)和图 3(d)分别为 SE-only、FE-only 和 DF-Former 所对应的 GradCAM 热力图,三者对图 3(a)的分类预测结果均正确。但是,GradCAM 热力图显示:FE-only 仅学习到了蜗牛壳体的纹理模式和整体形状,且误将蜗牛壳体上附着的细小砂石这一噪声特征进行了建模;SE-only 仅学习到了蜗牛软体和部分反光的壳体,而忽视了大部分壳体;而 DF-Former 完整地识别到了蜗牛的蜗牛软体和壳体,且没有对噪声建模。GradCAM 可视化结果进一步验证了 DF-Former 内在结构设计的合理性,通过将空域所提取的局部特征和频域所提取的区域全局特征进行融合,从而能够大幅提升模型的精度和可靠性。SE 模块能够有效分辨微小的局部细节,从而避免对噪声建模。同时,由于整体形状和纹理模式的分辨难度小于微小的细节,因此先降维后升维可以降低模型计算量,且不会损失区域和全局相关特征信息。这同时也表明,为区域和全局特征的提取分配大量计算资源是不合理的。在二倍下采样的条件下,通过在频域学习区域全局特征,可以有效避免自注意力机

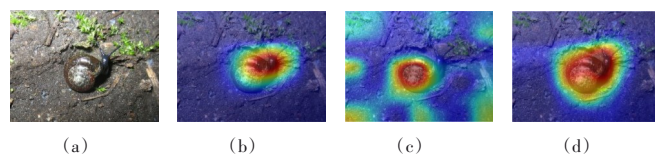


Fig. 3 GradCAM heatmap

图 3 GradCAM 热力图

制带来的高额计算开销。

4 结语

本文提出了以“分治—引导—融合”为核心的频率分支结构(SFF),其包含3个子模块:①空域编码器SE,用于在空域学习高频局部细节,其通过最大池化、部分卷积和空间移位等进行操作;②频域编码器FE,用于在频域学习区域和全局特征,其通过傅里叶变换将特征图映射至频域,结合“先降维再升维”策略提取区域和全局特征,显著降低了计算开销;③域特征融合模块DF,用于空域与频域特征的跨域融合,其通过逐点卷积和ECA注意力机制增强了模型对多尺度信息的表达能力。在SFF的基础上,本文进而构建一种新型的图像分类网络DF-Former。实验结果表明,在参数量和计算量低于ResNet50、Swin-Tiny、InceptionNeXt-T和IFormer-S的情况下,DF-Former在ImageNet100数据集上取得了与基准模型相近的性能。在低分辨率数据集CINIC-10上,DF-Former的优势则更显著,Top-1准确率分别提升了10.74%、12.2%、3.93%和7.71%。这一差异表明,DF-Former通过均衡分配局部与全局特征的计算资源,在保持轻量化架构的同时,对复杂视觉场景具有更强的泛化能力。

DF-Former在小型数据集CINIC-10上的各项指标均优于基准模型,但在大型数据集ImageNet-100上,其准确率相较于基准模型仍有差距。后续研究将继续优化结构,在减少参数量和计算量的同时,尽可能提升分类准确率。

参考文献:

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.
- [2] DOSOVITSKIY A, et al. An image is worth 16x16 words: transformers for image recognition at scale[DB/OL]. <https://arxiv.org/pdf/2010.11929>.
- [3] 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.
- [4] 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, 2021: 10012-10022.
- [5] SI C, YU W, ZHOU P, et al. Inception transformer[J]. Advances in Neural Information Processing Systems, 2022, 35: 23495-23509.
- [6] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11976-11986.
- [7] YU W, ZHOU P, YAN S, et al. Inceptionnext: when inception meets convnext[DB/OL]. <https://arxiv.org/pdf/2303.16900>.
- [8] WANG S, VELDHUIS R, BRUNE C, et al. What do neural networks learn in image classification? a frequency shortcut perspective[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 1433-1442.
- [9] BAI J, YUAN L, XIA S T, et al. Improving vision transformers by revisiting high-frequency components [DB/OL]. <https://arxiv.org/pdf/2204.00993>.
- [10] CHEN J, KAO S, HE H, et al. Run, don't walk: chasing higher flops for faster neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12021-12031.
- [11] RAO Y, ZHAO W, ZHU Z, et al. Global filter networks for image classification[J]. Advances in Neural Information Processing Systems, 2021, 34: 980-993.
- [12] WANG G, ZHAO Y, TANG C, et al. When shift operation meets vision transformer: an extremely simple alternative to attention mechanism[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 2423-2430.
- [13] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.
- [14] 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, 2009: 248-255.
- [15] DARLOW L N, et al. Cinic-10 is not imagenet or cifar-10 [DB/OL]. <https://arxiv.org/pdf/1810.03505>.

(责任编辑:陈维捷)