◆生物信息◆

MSViT:融合多尺度特征的轻量化图像分类混合模型*

覃　晓[1,2],彭　磊[1],廖惠仙[3],元昌安[4**],赵剑波[1],邓　超[1],钱泉梅[1],卢虹妃[1],龚远旭[1]
(1. 南宁师范大学广西人机交互与智能决策重点实验室,广西南宁　530100;2. 广西区域多源数据集成与智能处理协同创新中心,广西桂林　541004;3. 广东财贸职业学院数字技术学院,广东清远　511510;4. 广西科学院,广西南宁　530007)

摘要:针对现有 Vision Transformer (ViT) 模型在局部特征捕捉和多尺度特征融合方面的局限性,本文提出一种新型的融合多尺度特征的轻量化图像分类混合模型(Multi-Scale Vision Transformer,MSViT)。首先,在编码器中设计捕获通道特征的多尺度前馈神经网络(Multi-Scale Feed Forward Network,MSFFN)模块,该模块能有效提取空间和多尺度通道特征。其次,设计一个新的级联特征融合解码器(Cascade Feature Fusion Decoder,CFFD),通过整合特征金字塔网络(Feature Pyramid Network,FPN)和多阶段特征融合解码器,显著提升模型对不同尺度特征的交互和融合能力。最后,模型引入多阶损失函数,以全面优化不同尺度特征在图像分类任务中的表现。为了验证 MSViT 的有效性,在 4 个实验数据集[ImageNet-1k 的 1 个子集(Small_ImageNet)、Cifar 100、糖尿病视网膜病变数据集(APTOS 2019)、蘑菇数据集(Mushroom 66)]上进行大量的实验。其中在 Small_ImageNet 数据集上的实验结果显示,MSViT 实现了 87.58%的 Top-1 准确率,较 EdgeViT-XXS 提升了 2.27%。实验结果证明了 MSViT 在图像分类任务中的有效性。

关键词:图像分类;多尺度特征融合;多阶损失函数;特征金字塔网络(FPN);Transformer

中图分类号:TP391.41,TP183　　**文献标识码**:A　　**文章编号**:1005-9164(2024)05-0912-13
DOI:10.13656/j.cnki.gxkx.20241127.009

图像分类技术作为计算机视觉领域的基石,其在社会认知和实时应用中的重要性日益凸显。特别是轻量化图像分类模型,不仅能够减少计算和存储需求,而且能够在资源受限的移动环境中保持高效运行,这对于推动计算机视觉技术在移动设备上的广泛应用具有重要意义。在深度学习技术的推动下,卷积神经网络(Convolutional Neural Networks,CNN)[1]已成为图像分类领域的主流方法。CNN 使用卷积核在输入图像上提取局部特征,并通过逐层抽象构建起图像的高层表示。然而,CNN 的局部感受野限制了其捕捉图像全局上下文的能力,以及其在处理需要全面理解图像内容的任务中的表现。

收稿日期:2024-07-13　　修回日期:2024-10-14
*科技部科技创新 2030—"脑科学与类脑研究"重大项目(2021ZD0201904)和广西科技重大专项(桂科 AA22068057)资助。
【第一作者简介】
覃　晓(1973—),女,教授,主要从事数字图像处理、自然语言理解研究,E-mail:7670172@qq.com。
【**通信作者简介】
元昌安(1964—),男,博士,教授,主要从事智能计算研究,E-mail:yuanchangan@126.com。
【引用本文】
覃晓,彭磊,廖惠仙,等. MSViT:融合多尺度特征的轻量化图像分类混合模型[J]. 广西科学,2024,31(5):912-924.
QIN X,PENG L,LIAO H X,et al. MSViT:A Lightweight Image Classification Hybrid Model Integrating Multi-Scale Features [J]. Guangxi Sciences,2024,31(5):912-924.

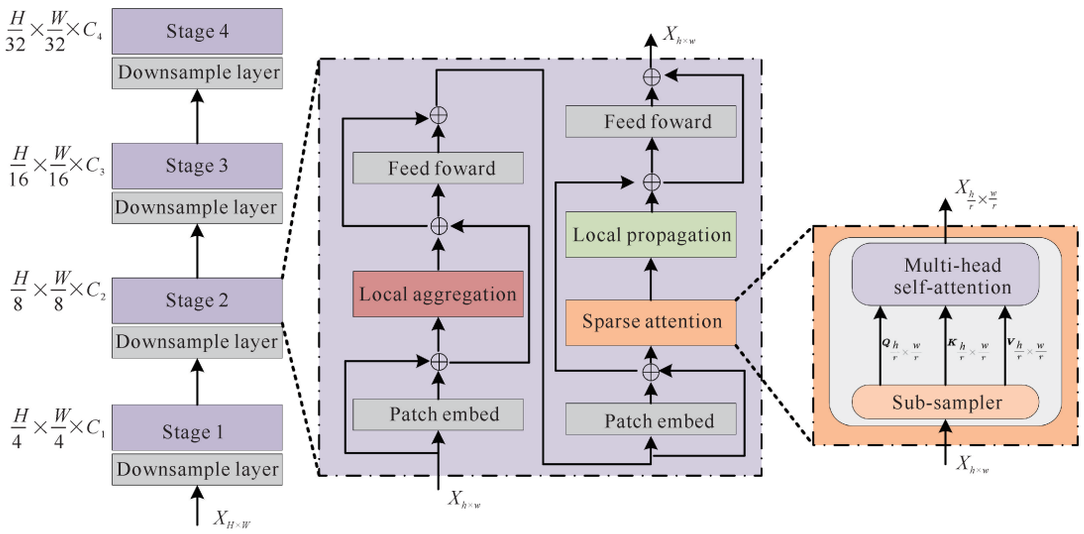
Dosovitskiy 等[2]提出的 Vision Transformer(ViT)首次将自然语言处理中极为成功的 Transformer 架构引入到计算机视觉任务中,在图像分类问题上取得了突破性进展。ViT 架构的核心组件是多头自注意力(Multi-Head Self Attention,MHSA)机制,它将特征图切分为多个固定大小的图像块(Patch)并进行独立编码,使得每个图像块都能够独立地参与到注意力计算中,通过多个自注意力头独立地计算注意力权重,最后对这些权重进行聚合,生成最终的全局特征表示。MHSA 机制使得 ViT 拥有强大的全局特征捕捉能力,但受到图像切分操作的干扰,MHSA 机制的局部特征提取能力存在天然缺陷。这一缺陷源于 MHSA 机制将特征图切分为多个独立的图像块后,仅在这些图像块之间进行注意力操作,而不会像 CNN 那样对图像块内部的特征进行卷积或者局部感受野的处理。此外,ViT 通常专注于单一尺度的学习,这也限制了其表达复杂视觉信息丰富性的能力,在运用到下游任务时,ViT 无法充分捕捉到关键的多尺度特征,从而导致多尺度信息的丢失。

可见,传统的 CNN 擅长提取图像的局部特征,但全局建模能力较弱,而 ViT 恰恰相反,它能够很好地捕捉全局特征,但提取局部特征和多尺度特征的能力不足。为了充分利用两者的优势,研究者们设计了一系列融合 ViT 和 CNN 优秀特性的混合模型,如 Pyramid Vision Transformer (PVT)[3]、PVT v2[4]、EdgeViT[5]、LeViT[6]、CMT[7]、LocalViT[8]、EfficientViT[9]、UNETR++[10]、SUNet[11]等。Wang 等[3]提出的轻量化混合模型 PVT 在 ViT 的基础上,集成了特征金字塔网络(Feature Pyramid Network,FPN)[12]结构。该结构的设计灵感来源于如 Resnet50[13]等经典模型,通过在 ResNet50 等模型的每个层次(Stage)递减特征图的空间尺寸,并相应地增加特征维度,实现对多尺度特征的提取。然而,PVT 采用的这一设计模式在实际应用中存在一定的局限性,具体而言,每个层次完成特征提取等任务后,该层次的特征信息便被丢弃,这种处理方式在一定程度上制约了模型在同一个层次内提取和利用多尺度信息的能力。另外,针对 ViT 中 MHSA 机制计算成本高的问题,PVT 引入了空间降维注意力(Spatial-Reduction Attention,SRA)机制。SRA 机制的核心在于,在执行注意力操作之前,先对特征图进行下采样,降低键(Key)和值(Value)的维度。SRA 机制在保证性能的同时,显著降低了模型的计算开销,使 PVT 在资源受限环境中具有更高的适用性。Pan 等[5]提出的 EdgeViT 也采用类似 PVT 的 FPN 结构来提取多尺度的特征信息。然而,PVT 和 EdgeViT 在执行最终的分类任务时,仅利用了模型中最后一个层次的特征信息。尽管以上模型都通过 FPN 结构成功提取了不同层次的特征,但这些特征并未通过解码等技术手段进行有效的多尺度融合,这极大地限制了模型在综合利用不同尺度特征方面的潜力。

针对上述问题,本文构建了融合多尺度特征的轻量化图像分类混合模型(Multi-Scale Vision Transformer,MSViT),MSViT 主要包含两个模块:融合空间和多尺度通道特征的编码器(Encoder Incorporating Spatial and Multi-scale Channel Features,ESC)模块、级联特征融合解码器(Cascade Feature Fusion Decoder,CFFD)模块。其中,ESC 模块融入多尺度前馈神经网络(Multi-Scale Feed Forward Network,MSFFN),可以有效提取多尺度通道特征信息,增强局部特征的表达;CFFD 模块引入 FPN 来加强多尺度特征的交互,并通过多层次聚合损失来增强图像分类精度,同时加速训练收敛。

1 相关工作

在探索 ViT 的多尺度特征提取表示能力的过程中,Pan 等[5]基于 ViT 提出一种轻量化的 FPN 模型 EdgeViT,并提出一种结合 CNN 局部特征提取优势和 ViT 全局特征捕捉能力的轻量化混合架构。EdgeViT 模型整体框架如图 1 所示。



(a) Architecture (b) Local-Global-Local Bottleneck (LGL) (c) Global Sparse Attention (GSA)

In this example, H and W represent the initial height and width of the input image of the model respectively, while C_i is the number of channels of layer i, and h and w represent the height and width of the feature map of the current layer (layer i), respectively. r is the sub-sampling rate.

图 1 EdgeViT 模型整体框架

Fig. 1 Overall frame of EdgeViT model

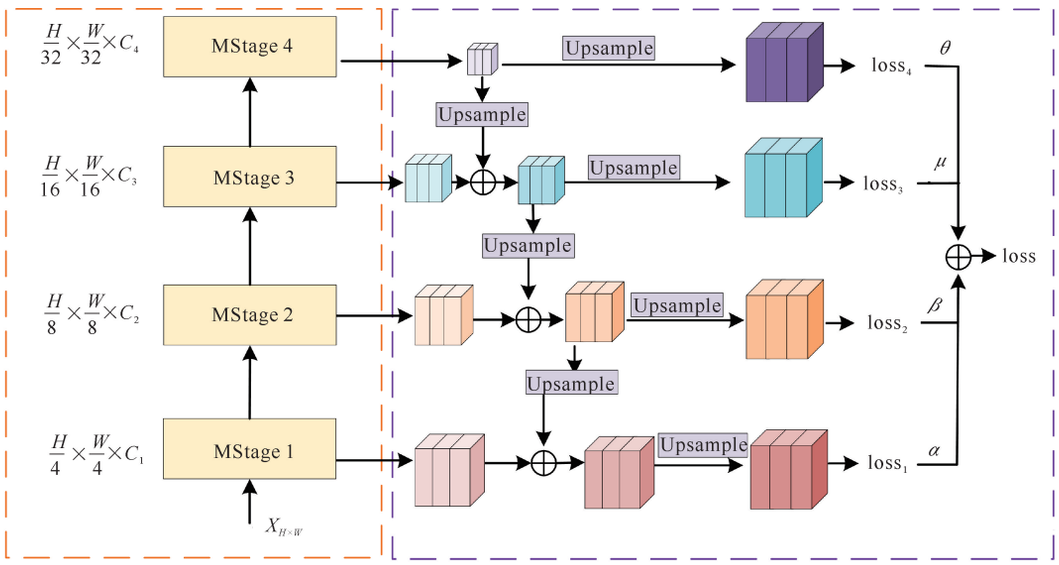
首先,EdgeViT 引入了 FPN 结构,能够实现在不同的层次上采用不同大小的 Transformer 块。这允许模型在早期层次使用较小、计算成本较低的块,而在更深的层次则使用较大、能力更强的块。这种层次化方法有助于平衡模型的性能及其计算成本,使得模型即使在资源有限的设备上也能有效运行。其次,在每个层次中,EdgeViT 引入了 Local-Global-Local Bottleneck(LGL)机制,这一机制的核心在于通过分阶段的注意力计算来平衡局部和全局特征的表示。LGL 机制的工作流程如下:首先在局部图像块(Token)之间进行局部特征聚合,获得一组有代表性的 token;然后将这一组 token 输入到全局稀疏注意力(Global Sparse Attention,GSA)中建模 token 之间的全局关系,并将学习到的上下文信息扩散回局部非代表 token 中;最后通过前馈神经网络(Feed Forward Network,FNN)进一步增强特征的表示能力。最终,模型根据最高层级的特征信息,即最后一个层次输出的特征,来执行分类任务。这一设计策略使得 EdgeViT 能够在减少计算复杂度的同时,有效地捕捉、融合局部和全局的视觉信息,为图像识别任务提供了一种新的视角。

尽管 EdgeViT 在多尺度和稀疏注意力方面有所创新,但是其特征融合过程仍有改进空间,特别是在实现不同尺度特征间的深度交互和融合方面。值得注意的是,在 LGL 模块的设计中,FNN 组件仅采用了多层感知器(Multi-Layer Perceptron,MLP)对特征图进行基础的特征提取,而并未充分考虑在不同尺度下特征信息的多样性和复杂性。本文提出的 MSViT 在 EdgeViT 的基础上,进一步探讨了如何改进 FNN 组件,以更有效地捕获和整合局部的多尺度信息。

2 方法

2.1 模型概述

MSViT 在基于 ViT 架构获取全局空间和通道信息的同时,又融合 FPN 结构以提取多尺度局部特征,因此能更好地处理图像分类任务。由图 2 可见,MSViT 主要由 ESC 模块和 CFFD 模块构成。其中 ESC 模块主要由 4 个层次(MStage)组成,每个 MStage 包含用于提取空间特征信息的 GSA 模块和捕获通道特征信息的 MSFFN 模块。ESC 具体实现方式是在基准编码器中融入 MSFFN 模块,提取多尺度通道特征信息,最后将空间特征和多尺度通道特征进行信息融合。CFFD 以级联的方式,将 FPN 结构融入 ViT 架构中,加强不同分辨率的特征交互。此外,从分层解码器的不同阶段聚合和优化多个损失,提升图像分类的精确度和训练收敛速度。



(a) ESC (b) CFFD

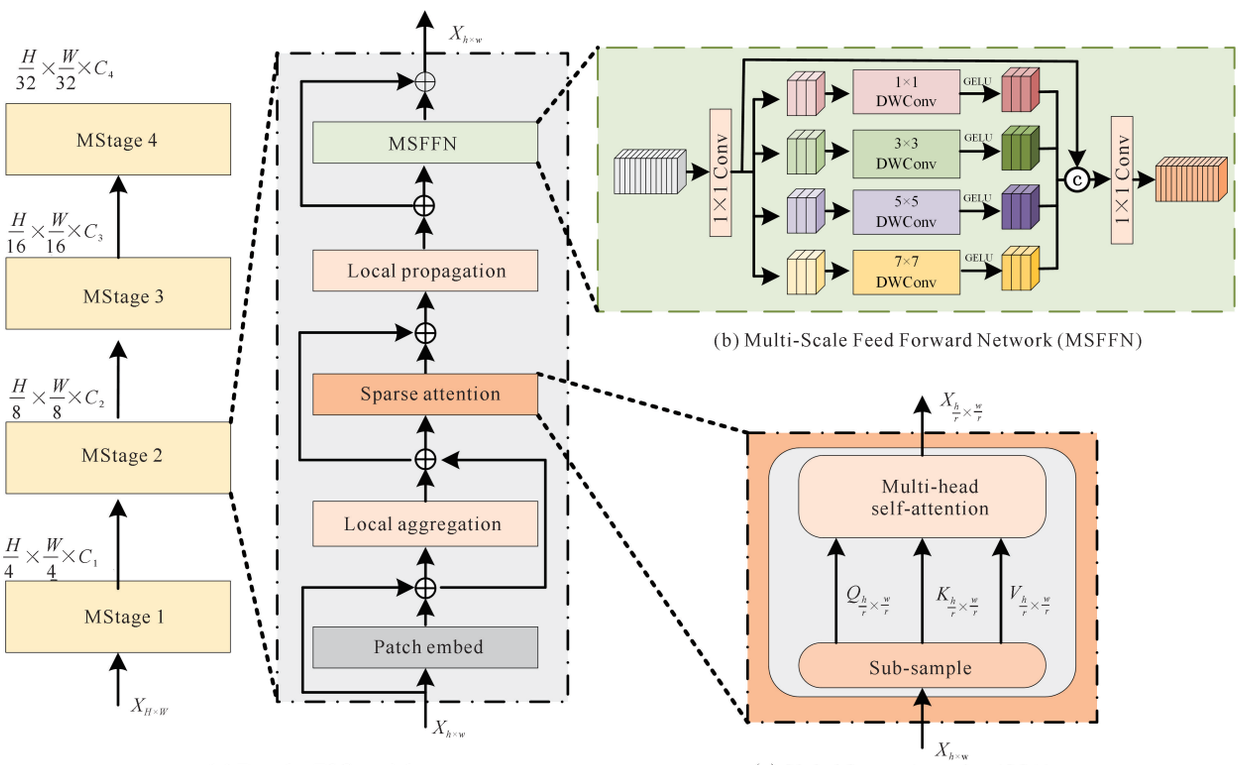
In this example, H and W represent the initial height and width of the input image of the model respectively, while C_i is the number of channels of layer i. θ, μ, β and α represent the loss weight of a single prediction head.

图 2 MSViT 架构

Fig. 2 Architecture of MSViT

2.2 融合空间和多尺度通道特征的编码器(ESC)

图 3 是 ESC 模块中每个 MStage 的详细结构图，其中本文提出的提取通道特征信息的 MSFFN 模块如图 3(b)所示。



(a) Encoder ESC module (b) Multi-Scale Feed Forward Network (MSFFN) (c) Global Sparse Attention (GSA)

In this example, H and W represent the initial height and width of the input image of the model respectively, while C_i is the number of channels of layer i, and h and w represent the height and width of the feature map of the current layer (layer i), respectively. r is the sub-sampling rate. Q is Query, K represents Key, V represents Value.

图 3 融合空间和多尺度通道特征的编码器结构图

Fig. 3 Structure diagram of encoder integrating spatial and multi-scale channel features

多尺度通道特征反映了图像在不同分辨率下的通道特征信息,而这些信息对图像分类任务非常重要。传统前馈神经网络在特征提取方面主要依赖于固定大小的输入,这限制了它们处理不同尺度特征的能力,难以充分利用多尺度通道特征。为此,模型将经过 GSA 模块处理后输入 MSFFN 模块的特征图按照通道维度(C)进行切分,得到 K 个特征图子集。随后对每个子集的特征图独立应用深度可分离卷积提取到多尺度的特征信息。这种方法不仅提高了模型对多尺度通道特征的利用效率,而且增强了其对不同尺度信息的捕捉能力。其具体步骤如下:首先在每个 MStage 中,对输入 MSFFN 模块的特征图 $X_S \in \mathbb{R}^{H\times W\times C}$,按照 C 平均切分得到 K 个特征图子集 $X_k \in \mathbb{R}^{H\times W\times CK}$,其中,$H$ 表示图像的高度(Height),W 表示图像的宽度(Width),$CK=C/K$;然后分别对每个特征图子集 $X_i \in \mathbb{R}^{H\times W\times C_i}$($i \in [1,K]$)使用 K 个大小不同的卷积核做深度可分离卷积,获取多尺度通道特征信息。同时,模型为了确保原始特征的完整性,除了应用 K 个不同尺寸的卷积核提取特征外,MSFFN 模块还引入了残块结构,弥补丢失的特征;最后,将 $K+1$ 个特征图子集按通道拼接,再经过一个逐点卷积,在深度方向上进行加权组合,有效整合通道的多尺度特征,增强模型对复杂数据结构的表征能力。MSFFN 模块的伪代码如下所示。

算法 1　MSFFN(X)
输入:$X_S \in \mathbb{R}^{H\times W\times C}$　//输入特征矩阵 X_S
输出:$X' \in \mathbb{R}^{H'\times W'\times C'}$　//输出特征矩阵 X'
{
$X_C=\mathrm{Conv}(1\times 1,X_S)$;
$(X_1\cdots X_k)=\mathrm{split}(X_C,\mathrm{dim})$;　//按通道切分成 k 份
for($i=1$ to k)
$X_i=\mathrm{GELU}(\mathrm{DWConv}((2i-1)\times(2i-1),X_i))$; //执行不同卷积核的深度可分离卷积
$X_C=\mathrm{concat}(X_C,X_i)$;
end for
$X'=\mathrm{Conv}(1\times 1,X_C)$;　//逐点卷积,深度方向加权组合特征信息
$X'=X'+X_S$;　//空间特征和多尺度通道特征融合
return X';
}

2.3　级联特征融合解码器(CFFD)

传统的 EdgeViT 利用最后一个层次获得的特征进行图像分类。然而在模型运行期间,每一个层次的执行都获得了不同尺度、不同空间位置的局部信息,为了充分利用这些信息以取得更好的分类效果,本文提出一种新的 CFFD。如图 2(b)所示,CFFD 主要由特征融合和多阶损失优化两个部分组成。

2.3.1　特征融合

记每个 MStage 的输出特征图为 MStage[i],为了匹配 MStage[$i-1$]阶段的特征并进行融合,将 MStage[i]输出的特征图做上采样操作,与 MStage[$i-1$]的输出进行特征融合,最后计算 MStage[i]的预测概率 y_i,具体操作如公式(1)和(2)所示,其中,Upconv()为上采样函数。

$$\mathrm{MStage_{up}}[i]=\mathrm{Upconv}(\mathrm{MStage}[i+1])+\mathrm{MStage}[i], \tag{1}$$

$$y_i=\mathrm{Softmax}(\mathrm{MStage_{up}}[i]), \tag{2}$$

2.3.2　多阶损失优化

为了聚合和优化图像最终的分类性能,每个 MStage 均单独计算 1 个损失,最后 4 个损失累加算出最终预测结果。本文采用交叉熵函数计算每个图像的真实概率 p 和预测概率 y_i 之间的损失,具体计算公式如(3)和(4)所示:

$$\mathrm{loss}_i(p,y_i)=-\sum_{i=1}^{n} p\cdot\log(y_i), \tag{3}$$

$$\mathrm{loss}=\alpha\cdot\mathrm{loss}_1+\beta\cdot\mathrm{loss}_2+\mu\cdot\mathrm{loss}_3+\theta\cdot\mathrm{loss}_4, \tag{4}$$

其中,n 代表类别数量,loss_i 代表第 i 个 MStage 预测头的损失,α、β、μ 和 θ 是单个预测头的损失权重,在实验中,α、β、μ 和 θ 均设置为 0.25。根据 4 个预测头计算总损失 loss,模型能够动态地调整图像分类的最优预测并输出最终预测概率,从而有效地提升模型性能。CFFD 模块的伪代码如下所示。

算法 2　CFFD(X)
输入:$X \in \mathbb{R}^{H\times W\times C}$　//输入特征矩阵 X
输出:预测概率 y　//输出预测概率
{
for $i=1$ to n:　//循环调用算法 1 得到 4 个 MStage 的输出
$X_S=\mathrm{GSA}(X)$;　//GSA(X) 为执行全局稀疏注意力函数
MStage[i]=MSFFN(X_S);
end for
$\mathrm{MStage_{new}}[n+1]=\mathrm{MStage}[n]$;

for $i=n$ to 1:　　//循环计算第 4 至第 1 个 stage 的 loss

$\text{MStage}_{new}[i]=\text{MStage}_{new}[i+1]+\text{MStage}[i]$;　//特征融合

$\text{MStage}_{up}[i]=\text{Upconv}(\text{MStage}_{new}[i])$;　//上采样操作

$y_i=\text{Softmax}(\text{MStage}_{up}[i])$;　//求第 i 层预测概率

$\text{loss}_i(p,y_i)=-\sum_{i=1}^{n}p\cdot\log(y_i)$;　//求第 i 层 loss

end for

$\text{loss}=\alpha*\text{loss}_1+\beta*\text{loss}_2+\mu*\text{loss}_3+\theta*\text{loss}_4$;　//多阶损失优化

return 最优预测概率 y;

}

3　实验与结果分析

3.1　数据集

为了验证本文提出的 MSViT 的有效性和泛化性,在 4 个图像分类数据集上进行实验,分别为 ImageNet-1k 的 1 个子集(Small_ImageNet)、Cifar 100、糖尿病视网膜病变数据集(APTOS 2019)、蘑菇数据集(Mushroom 66),这 4 个数据集的图像数量如表 1 所示。

表 1　4 个数据集的图像数据信息

Table 1　Image data information for four datasets

数据集 Dataset	训练数量 Training quantity	验证数量 Verified quantity	类别数 Number of classes
Small_ImageNet	80 000	20 000	1 000
Cifar 100	50 000	10 000	100
APTOS 2019	8 000	2 000	5
Mushroom 66	12 757	3 190	66

Small_ImageNet 是 Kaggle(www.kaggle.com)上一个公开的数据集,由 DeepMind 团队创建,用于支持资源受限情况下的图像识别研究。它包含 1 000 个类别,每个类别 100 张图片,共计 10 万张图片。由于硬件限制,选择了 Small_ImageNet 作为实验数据集,它特别适用于小样本学习(Few-shot learning)的研究。通过使用该数据集,研究人员可以在有限的资源下探索和验证模型的性能。

Cifar 100 是图像分类中最常用的数据集之一,共有 6 万张照片,主要包含 100 个类别,每个类别有 600 张照片。其中 20 个高级分类,每个高级分类里面又包含 5 个子类。相较于 Cifar 10 数据集,Cifar 100 数据集更具挑战性,常用于细粒度分类任务。

APTOS 2019 也是 Kaggle 上的公开数据集,是由印度的 Aravind 眼科医院收集的糖尿病视网膜病变图构成,可用于眼科疾病的检测研究,该数据集共有 5 个类别,每个类别有 2 000 张照片。

Mushroom 66 是本团队在专业人员指导下收集所得的私有数据集,该数据集共有 15 947 张照片,共有 66 个类别。Mushroom 66 数据集的图像大部分是在自然场景下的蘑菇图片,背景噪声较多,符合实际应用场景。

3.2　实验设备和超参设置

实验环境为 Ubuntu 18.04.6 LTS 系统,CPU 为 Intel Core i9-10980XE CPU@3.00GHz,显卡 GPU 为 NVIDIA GeForce RTX4080,显存 16 GB。实验基于 PyTorch 深度学习框架,开发环境为 PyTorch 1.10.1,Cuda 11.8,Python 3.8。模型训练迭代 200 个 epochs,训练批次大小为 64,初始学习率为 0.001,weight_decay 为 0.000 1,优化器使用 Adam。

3.3　实验说明

为了保证实验结果的公平性,所有模型均不使用预训练权重,且图片输入统一为 224×224,所有模型的训练参数设置保持一致。实验首先对 Small_ImageNet、APTOS 2019 和 Mushroom 66 3 个数据集以训练集∶验证集=8∶2 的比例进行随机划分,这一比例是参考了当前主流模型[4,5,14]的实验设置。对于 Cifar 100 数据集,则是遵循官方的分割比例,即训练集∶验证集=5∶1,这是根据该数据集的标准实践所得。其次,为了确保实验的可复现性,在实验中特别设置了随机数种子,以增强对比实验的可靠性和可对比性。

3.4　对比实验

在相同实验环境下,将本文提出的 MSViT 和当前主流的图像分类模型在未使用预训练权重的前提下进行详细实验和数据对比。对比模型有 RepMLP-Net-T[15]、Hornet-$T_{7\times7}$[16]、FasterNet-T[17]、TransX-Net-T[18]、Agent-PVT-T[19]、RepVGG-A0[20]、MPViT-T[21]、PVT v2-B0[4]、MobileViT-XS[22] 和 EdgeViT-XXS[5]等,实验指标包含参数量(Params)、浮点计算量(FLOPs)、Top-1 准确率和 Top-5 准确率,结果详见表 2 至表 5(其中加粗数据表示各模型中的最高准确率)。

表 2　各模型在 Small_ImageNet 数据集中的实验结果

Table 2　Experimental results of each model on the Small_ImageNet dataset

模型 Model	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
RepMLPNet-T[15](2022)	31.39	1.09	84.85	95.25
Hornet-$T_{7\times7}$[16](2022)	22.41	3.99	84.67	95.03
FasterNet-T[17](2023)	14.98	1.91	61.32	84.23
TransXNet-T[18](2023)	12.83	1.77	83.18	94.73
Agent-PVT-T[19](2023)	11.57	1.93	83.78	94.80
RepVGG-A0[20](2021)	9.11	1.52	86.61	96.31
MpViT-T[21](2022)	5.84	1.83	86.14	96.31
PVT v2-B0[4](2022)	3.44	0.53	72.46	89.92
MobileViT-XS[21](2021)	2.32	0.72	79.60	93.62
EdgeViT-XXS[5](2022)	4.07	0.54	85.31	95.50
MSViT (ours)	4.95	0.81	**87.58**	**96.33**

表 3　各模型在 Cifar 100 数据集中的实验结果

Table 3　Experimental results of each model on the Cifar 100 dataset

模型 Model	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
RepMLPNet-T[15](2022)	31.39	1.09	45.51	73.84
Hornet-$T_{7\times7}$[16](2022)	22.41	3.99	47.24	71.54
FasterNet-T[17](2023)	14.98	1.91	58.04	83.64
TransXNet-T[18](2023)	12.83	1.77	64.08	84.54
Agent-PVT-T[19](2023)	11.57	1.93	60.12	83.89
RepVGG-A0[20](2021)	9.11	1.52	68.88	88.47
MpViT-T[21](2022)	5.84	1.83	68.15	89.78
PVT v2-B0[4](2022)	3.44	0.53	66.36	86.82
MobileViT-XS[22](2021)	2.32	0.72	68.82	90.81
EdgeViT-XXS[5](2022)	4.07	0.54	70.88	90.19
MSViT (ours)	4.95	0.81	**73.18**	**90.96**

表 4　各模型在 APTOS 2019 数据集中的实验结果

Table 4　Experimental results of each model on the APTOS 2019 dataset

模型 Model	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
RepMLPNet-T[15](2022)	31.39	1.09	79.40	99.60
Hornet-$T_{7\times7}$[16](2022)	22.41	3.99	71.30	99.90
FasterNet-T[17](2023)	14.98	1.91	67.65	99.90
TransXNet-T[18](2023)	12.83	1.77	74.95	99.95
Agent-PVT-T[19](2023)	11.57	1.93	73.45	**100.00**

续表

Continued table

模型 Model	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
RepVGG-A0[20](2021)	9.11	1.52	83.00	99.90
MpViT-T[21](2022)	5.84	1.83	76.05	99.80
PVT v2-B0[4](2022)	3.44	0.53	62.20	**100.00**
MobileViT-XS[22](2021)	2.32	0.72	62.30	**100.00**
EdgeViT-XXS[5](2022)	4.07	0.54	72.85	**100.00**
MSViT (ours)	4.95	0.81	**83.90**	99.90

表 5　各模型在 Mushroom 66 数据集中的实验结果

Table 5　Experimental results of each model on the Mushroom 66 dataset

模型 Model	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
RepMLPNet-T[15](2022)	31.39	1.09	78.20	93.93
Hornet-$T_{7\times7}$[16](2022)	22.41	3.99	68.50	92.86
FasterNet-T[17](2023)	14.98	1.91	60.20	86.03
TransXNet-T[18](2023)	12.83	1.77	78.77	93.68
Agent-PVT-T[19](2023)	11.57	1.93	73.91	92.73
RepVGG-A0[20](2021)	9.11	1.52	81.36	95.38
MpViT-T[21](2022)	5.84	1.83	82.62	95.26
PVT v2-B0[4](2022)	3.44	0.53	72.01	91.72
MobileViT-XS[22](2021)	2.32	0.72	82.50	95.86
EdgeViT-XXS[5](2022)	4.07	0.54	78.33	94.18
MSViT (ours)	4.95	0.81	**83.20**	**95.89**

表 2 是 MSViT 与当前主流图像分类模型在 Small_ImageNet 数据集上的比较分析结果。本实验中,MSViT 在 Small_ImageNet 数据集上的表现突出,其 Top-1 准确率达到了 87.58%,相较于基线模型 EdgeViT-XXS 提升了 2.27%。此外,MSViT 的 Top-5 准确率相较于基线模型 EdgeViT-XXS 也实现了接近 1%的提升,与其他主流图像分类模型相比,MSViT 取得了更高的准确率。Small_ImageNet 数据集包含了 1 000 个类别,每个类别在尺寸和形态上存在差异,这对模型的泛化和适应性提出了更高要求。MSViT 通过 MSFFN 模块可以有效增强多尺度特征的表达能力,进而全面关注每个类别的特征。此外,模型中的 CFFD 模块进一步优化了特征整合,辅以多损失函数的联合优化策略,有效提升了预测精度。这些设计使得 MSViT 在 Small_ImageNet 数据集上取得了优异的表现。

同理,在 Cifar 100、APTOS 2019 以及 Mushroom 66 这 3 个数据集上深入探讨了 MSViT 的性能并与其他模型进行了详尽的对比实验。这些实验旨在全面评估 MSViT 在不同数据集上的表现,并与现有的先进模型进行比较。表 3、表 4 和表 5 分别展示了 MSViT 在 Cifar 100、APTOS 2019 和 Mushroom 66 数据集上得到的实验结果,这些结果不仅提供了 MSViT 性能的直观展示,而且也为进一步的模型优化和算法改进提供了重要的参考依据。这 3 个数据集各具特性:在 Cifar 100 数据集中,超类中子类之间存在高度的相似性,这要求模型能够区分细微尺度的特征差异;APTOS 2019 数据集包含了视网膜病变的早期迹象,如血管壁变薄导致的深红色点状物和暗斑,以及毛细血管坏死时的亮点絮状物或白斑,这些特征要求模型具备高度的敏感性;Mushroom 66 数据集的自然场景图像具有复杂的背景、不同的图像尺寸以及目标物体的遮挡、光照和模糊等变量,这对模型的特征提取能力提出了更高的要求。MSViT 针对

这些挑战,通过在不同尺度上观察并抽取图像的关键信息并将特征交互融合,有效地整合了多尺度特征,让模型具备更强大的分类性能,进而在各个数据集上都取得了优异的性能表现。

根据以上实验结果可知,本文提出的 MSViT 在所有数据集上均表现出色。MSViT 具有更好的分类效果主要有以下原因:首先,相较于 RepMLPNet-T、RepVGG-A0 等纯卷积深度学习模型,MSViT 融入了 MHSA 机制,能够更好地提取全局的特征信息,增强了全局特征的表达能力;其次,与 Hornet-$T_{7\times7}$、TransXNet-T、Agent-PVT-T、FasterNet-T、MpViT-T 和 EdgeViT-XXS 等模型相比,MSViT 在全连接层融入了多尺度特征增强模块,有针对性地捕获和增强多尺度特征信息,更有利于图像分类任务;最后,MSViT 不仅在 FNN 中融入多尺度信息,还通过级联的方式将多尺度特征进行交互融合,加深了模型的特征表达,进而显著提升了模型性能。

为了更直观地观察对比实验中主要模型在 4 个数据集上的表现,将 4 个数据集的 Top-1 准确率实验结果通过柱状图进行展示,如图 4 所示。

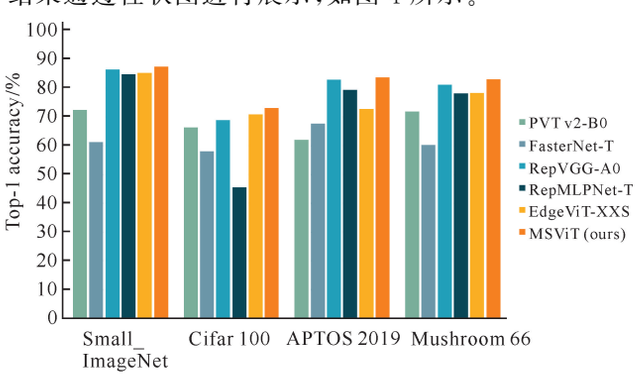


图 4　主要模型在 4 个数据集上的 Top-1 准确率

Fig. 4　Top-1 accuracy of the main model on the four datasets

从表 2 至表 5 以及图 4 的柱状图可以看出,MSViT 在 Top-1 和 Top-5 准确率上取得了显著提升。模型的准确率是一个关键的性能指标,但模型的轻量化程度也是研究中不可忽视的一个重要方面。如图 5 所示,通过对比主要模型(包括 EdgeViT-XXS 等)的 Params 和 FLOPs 指标,可以发现 MSViT 在提高 Top-1 和 Top-5 准确率的同时,相较于 FasterNet、RepVGG-A0、RepMLP-T 等模型,其 Params 和 FLOPs 均有不同程度的减少。但与基准轻量化混合模型 EdgeViT-XXS 相比,MSViT 的 Params 和 FLOPs 却有一定的增加量。具体而言,MSViT 通过引入 MSFFN 和 CFFD 模块,导致 Params 增加了 0.88 M,FLOPs 增加了 0.27 G。尽管如此,Params 和 FLOPs 的增加带来了显著的性能提升:MSViT 的 Top-1 准确率在 Small_ImageNet、Cifar 100、APTOS 2019 和 Mushroom 66 数据集上分别提升了 2.27%、2.30%、11.05%和 4.87%。因此,本文认为这种参数量和计算量的增加是合理的,并且与所获得的性能提升相匹配。

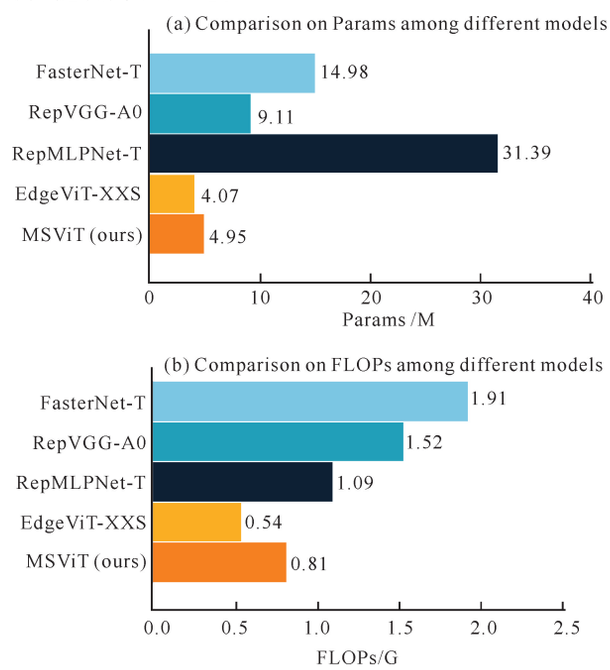


图 5　不同模型 Params 与 FLOPs 指标对比

Fig. 5　Comparison on Params and FLOPs among different models

3.5　消融实验

本文提出的 MSViT 主要包含 MSFFN 和 CFFD 两大模块。为了验证本文提出模块的有效性,本节在 4 个数据集上分别对 MSFFN 和 CFFD 两个模块进行消融实验。消融实验依旧以 Params、FLOPs、Top-1 准确率和 Top-5 准确率为评估指标。以 EdgeViT-XXS 为基准模型,在 4 个数据集上的消融实验结果如表 6 至表 9 所示。其中,√为使用该模块,×为不使用该模块;加粗数据表示各模型中的最高准确率。

表 6　在 Small_ImageNet 数据集上的消融实验结果

Table 6　Results of ablation experiments on the Small_ImageNet dataset

模型 Model	MSFFN	CFFD	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
EdgeViT-XXS[5]	×	×	4.07	0.54	85.31	95.50
MSViT-1	√	×	4.17	0.56	86.69	96.20
MSViT-2	×	√	4.81	0.76	86.51	96.32
MSViT (ours)	√	√	4.95	0.81	**87.58**	**96.41**

表 7　在 Cifar 100 数据集上的消融实验结果

Table 7　Results of ablation experiments on the Cifar 100 dataset

模型 Model	MSFFN	CFFD	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
EdgeViT-XXS[5]	×	×	4.07	0.54	70.88	90.19
MSViT-1	√	×	4.17	0.56	72.19	90.24
MSViT-2	×	√	4.81	0.76	71.39	89.54
MSViT (ours)	√	√	4.95	0.81	**73.18**	**90.96**

表 8　在 APTOS 2019 数据集上的消融实验结果

Table 8　Results of ablation experiments on the APTOS 2019 dataset

模型 Model	MSFFN	CFFD	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
EdgeViT-XXS[5]	×	×	4.07	0.54	72.85	**100.00**
MSViT-1	√	×	4.17	0.56	79.10	99.90
MSViT-2	×	√	4.81	0.76	78.60	99.80
MSViT (ours)	√	√	4.95	0.81	**83.90**	99.90

表 9　在 Mushroom 66 数据集上的消融实验结果

Table 9　Results of ablation experiments on the Mushroom 66 dataset

模型 Model	MSFFN	CFFD	参数量/M Params/M	浮点计算量/G FLOPs/G	Top-1 准确率/% Top-1 accuracy/%	Top-5 准确率/% Top-5 accuracy/%
EdgeViT-XXS[5]	×	×	4.07	0.54	78.33	94.18
MSViT-1	√	×	4.17	0.56	81.81	95.01
MSViT-2	×	√	4.81	0.76	80.73	95.32
MSViT (ours)	√	√	4.95	0.81	**83.20**	**95.89**

图 6 为 MSViT 在不同数据集上的消融实验折线图。由表 6 和图 6(a)可知,在 Small_ImageNet 数据集上,在基准模型的基础上使用 MSFFN 模块,Top-1 准确率提升了 1.38%,说明提取多尺度特征非常必要;增加 CFFD 模块后,Top-1 准确率也提升了 1.20%,证明不同尺度的特征融合也有助于图像精准分类。既使用 MSFFN 模块,又使用 CFFD 模块时,实验效果最佳,Top-1 准确率能达到 87.58%,相比基准模型提升了 2.27%,Top-5 准确率达到了 96.41%,相比基准模型提升将近 1%。由表 7 和图 6(b)可知,在 Cifar 100 数据集上,在基准模型基础上使用 MSFFN 模块,Top-1 准确率提升了 1.31%;增加 CFFD 模块后,Top-1 准确率也有所提升,二者同时使用时,实验效果最佳,Top-1 准确率能达到

73.18%,相比基准模型提升了 2.30%。由表 8 和图 6(c)可知,在 APTOS 2019 数据集上,在基准模型基础上使用 MSFFN 模块,Top-1 准确率提升了 6.25%;增加 CFFD 模块后,Top-1 准确率提升了 5.75%,二者同时使用时,Top-1 准确率能达到 83.90%,相比基准模型提升了 11.05%,说明 MSViT 在实际应用数据集上效果更好。由表 9 和图 6(d)可知,在 Mushroom 66 数据集上,在基准模型的基础上引入 MSFFN 模块,模型的 Top-1 准确率显著提升了 3.48%。进一步地,当加入 CFFD 模块时,Top-1 准确率再次获得了 2.40%的提升。当这两种技术联合使用时,模型的 Top-1 准确率达到了 83.20%,与基准模型相比,实现了近 5%的显著提升。这一结果充分证明了模型在性能上的显著改进和有效性。

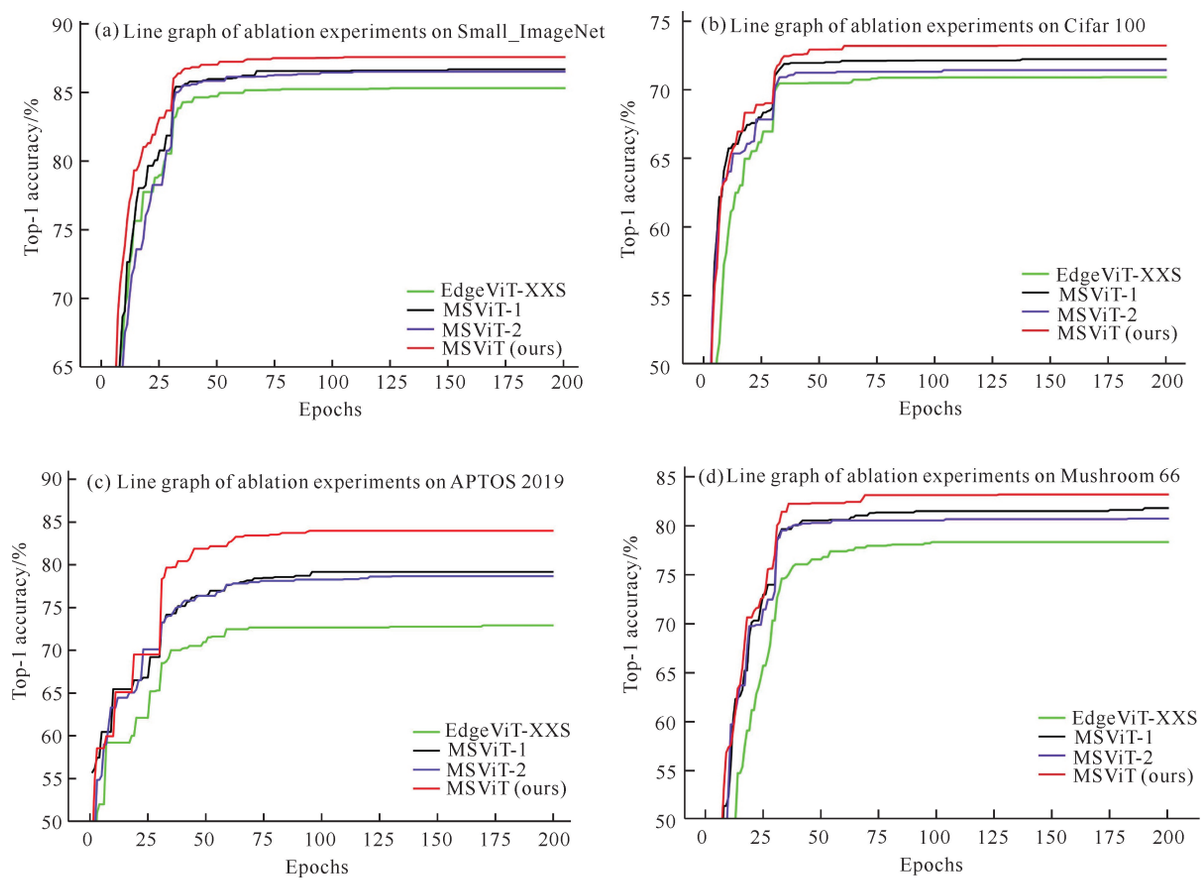


图 6　MSViT 在不同数据集上的消融实验折线图

Fig. 6　Line graphs of MSViT ablation experiments on different datasets

从以上 4 个数据集的消融实验结果可以观察到,无论是只添加 MSFFN 或 CFFD 模块,模型的图像分类性能均有所提升,尤其是在应用型数据集上提升效果非常明显,在 APTOS 2019 数据集上,二者同时使用时,Top-1 准确率提升了 11%,在 Mushroom 66 数据集上 Top-1 准确率提升了将近 5%。只添加 MSFFN 模块时,增加的 Params 和 FLOPs 几乎可以忽略不计,这是因为 MSFFN 模块的多尺度卷积使用的是深度可分离卷积。只添加 CFFD 模块时,Params 增加了 0.74 M,FLOPs 增加了 0.22 G。这主要是由于融入多尺度分层金字塔结构,层与层之间的矩阵运算会导致 Params 和 FLOPs 增加。

以上消融实验结果表明,本文所提出的 MSFFN 和 CFFD 模块均显著增强了图像分类的性能。这些结果不仅验证了 MSViT 在图像分类任务中的有效性,还证实了其在实际应用中的可靠性。

4　结束语

本文提出一种融合多尺度特征的轻量化图像分类混合模型(MSViT),该模型在基于 ViT 架构获取全局空间和通道信息的同时,又融合 FPN 结构提取多尺度局部特征,能更好地处理图像分类任务。MSViT 的整体架构由 ESC 和 CFFD 构成。其中 ESC 通过在 FNN 模块中整合多尺度卷积,有效增强了模

型对多尺度特征信息的捕捉能力。而 CFFD 利用级联机制实现多尺度特征的交互融合,显著提升了图像分类的性能。为了验证 MSViT 的可靠性和有效性,在 4 个不同的图像分类数据集上进行了广泛的对比实验和消融实验,实验结果一致表明 MSViT 在保持轻量化的同时,实现了对现有轻量化模型准确率的提升。

然而,通过分析实验结果,发现在要求细粒度特征提取能力较强的数据集上,MSViT 的性能提升幅度相对较小,模型在处理需要更精细特征识别的任务时还有可改进的空间。下一步,将深入探索如何进一步提升模型的细粒度特征提取能力,以实现在更广泛的图像分类任务中获得更均衡的性能提升。

参考文献

[1] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet classification with deep convolutional neural networks [C]//PEREIRA F,BURGES C J C,BOTTOU L,et al. Advances in Neural Information Processing Systems 25 (NIPS 2012). NY: Curran Associates, Inc., 2012,25:1097-1105.

[2] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al. An image is worth 16×16 words: transformers for image recognition at scale [Z/OL]. (2020-10-22)[2024-07-13]. https://doi.org/10.48550/arXiv.2010.11929.

[3] WANG W H,XIE E,LI X,et al. Pyramid vision transformer:a versatile backbone for dense prediction without convolutions [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE,2021:568-578.

[4] WANG W H,XIE E Z,LI X,et al. PVT v2:improved baselines with pyramid vision transformer [J]. Computational Visual Media,2022,8(3):415-424.

[5] PAN J,BULAT A,TAN F,et al. EdgeViTs:competing light-weight cnns on mobile devices with vision transformers [C]//Computer Vision－ECCV 2022. Lecture Notes in Computer Science,vol 13671. Cham:Springer,2022:294-311.

[6] GRAHAM B,EL-NOUBY A,TOUVRON H,et al. LeViT:a vision transformer in ConvNet's Clothing for faster inference [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE,2021:12239-12249.

[7] GUO J Y,HAN K,WU H,et al. CMT:convolutional neural networks meet vision transformers [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2022:12175-12185.

[8] LI Y W,ZHANG K,CAO J Z,et al. LocalViT:bringing locality to vision transformers [Z/OL]. (2021-4-12)[2024-07-13]. https://doi.org/10.48550/arXiv.2104.05707.

[9] LIU X Y,PENG H W,ZHENG N X,et al. EfficientViT:memory efficient vision transformer with cascaded group attention [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2023:14420-14430.

[10] SHAKER A,MAAZ M,RASHEED H,et al. UNETR++:delving into efficient and accurate 3D medical image segmentation [J]. IEEE Transactions on Medical Imaging,2024,43(9):3377-3390.

[11] FAN C M,LIU T J,LIU K H. SUNet:swin transformer UNet for image denoising [C]// Proceedings of the 2022 IEEE International Symposium on Circuits and Systems (ISCAS). Piscataway,NJ:IEEE,2022:2333-2337.

[12] LIN T Y,DOLLÁR P,GIRSHICK R,et al. Feature pyramid networks for object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2017:936-944.

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

[14] TOUVRON H,CORD M,DOUZE M,et al. Training data-efficient image transformers & distillation through attention [C]//Proceedings of the 38th International Conference on Machine Learning,PMLR 139,[S.l.:s.n],2021:10347-10357.

[15] DING X H,CHEN H H,ZHANG X Y,et al. RepMLPNet:hierarchical vision MLP with re-parameterized locality [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2022:578-587.

[16] RAO Y M,ZHAO W L,TANG Y S,et al. Hornet:efficient high-order spatial interactions with recursive gated convolutions [C]//Advances in Neural Information Processing Systems 35,[S.l.:s.n],2022:10353-10366.

[17] CHEN J R,KAO S H,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 (CVPR). Piscataway,NJ:IEEE ,2023:12021-12031.

[18] LOU M,ZHOU H Y,YANG S B,et al. TransXNet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition [Z/OL]. (2023-10-30)[2024-07-13]. https://doi. org/10.48550/arXiv. 2310. 19380.

[19] HAN D C,YE T Z,HAN Y Z,et al. Agent attention: on the integration of softmax and linear attention [Z/OL]. (2023-12-24)[2024-07-13]. https://doi. org/10.48550/arXiv. 2312. 08874.

[20] DING X H,ZHANG X Y,MA N N,et al. RepVGG: making VGG-style ConvNets great again [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2021:13728-13737.

[21] LEE Y,KIM J,WILLETTE J,et al. MPViT:multi-path vision transformer for dense prediction [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2022:7287-7296.

[22] MEHTA S,RASTEGARI M. MobileViT:lightweight,general-purpose,and mobile-friendly vision transformer [Z/OL]. (2021-10-5)[2024-07-13]. https://doi. org/10.48550/arXiv. 2110. 02178.

MSViT: A Lightweight Image Classification Hybrid Model Integrating Multi-Scale Features

QIN Xiao[1,2],PENG Lei[1],LIAO Huixian[3],YUAN Chang′an[4**],ZHAO Jianbo[1],DENG Chao[1],QIAN Quanmei[1],LU Hongfei[1],GONG Yuanxu[1]

(1. Guangxi Key Laboratory of Human-Computer Interaction and Intelligent Decision Making,Nanning Normal University,Nanning,Guangxi,530100,China;2. Guangxi Regional Collaborative Innovation Center for Multi-Source Data Integration and Intelligent Processing,Guilin,Guangxi,541004,China;3. College of Digital Technology,Guangdong Vocational College of Finance and Trade,Qingyuan,Guangdong,511510,China;4. Guangxi Academy of Sciences,Nanning,Guangxi,530007,China)

Abstract:Aiming at the limitations of existing Vison Transformer (ViT) models in local feature capture and multi-scale feature fusion,a new lightweight image classification hybrid model integrating multi-scale features (Multi-Scale Vision Transformer,MSViT) is proposed. Firstly,a Multi-Scale Feed Forward Network (MS-FFN) module is designed to capture channel features in the encoder,which can effectively extract spatial and multi-scale channel features. Secondly,a new Cascade Feature Fusion Decoder (CFFD) is designed. By integration of the Feature Pyramid Network (FPN) and the multi-stage feature fusion decoder,the interaction and fusion ability of the model for different scale features are significantly improved. Finally,a multi-order loss function is introduced to optimize the performance of different scale features in image classification tasks. To validate the effectiveness of the MSViT model,a large number of experiments are conducted on 4 datasets [a subset of ImageNet-1k (Small_ImageNet),Cifar 100,APTOS 2019,and Mushroom 66]. The experimental results on Small_ImageNet show that MSViT achieves the Top-1 accuracy of 87.58%,which is 2.27% higher than that of EdgeViTs-XXS. The experimental results demonstrate the effectiveness of MSViT in image classification tasks.

Key words:image classification;multi-scale feature fusion;multi-order loss function;Feature Pyramid Network (FPN);Transformer

责任编辑:梁　晓