文章编号:2096-1472(2024)01-0019-04
DOI:10.19644/j.cnki.issn2096-1472.2024.001.005

一种改进的 Swin Transformer 图像分类识别方法

陈 成[1],耿晓中[2],刘柏进[1],汪林恩[1],户唯新[2]

(1.吉林化工学院信息与控制工程学院,吉林 吉林 132022;
2.长春工程学院计算机学院,吉林 长春 130012)

✉ 2468295244@qq.com; dq_gxz@ccit.edu.cn; 1692797200@qq.com; 3172876826@qq.com; 1051090429@qq.com

摘 要:针对 Transformer 模型在处理图像任务时存在计算复杂度过大的问题,提出了一种改进的 Swin Transformer 图像分类识别方法。首先,Swin Transformer 使用补丁(Patch)化的图像特征图处理方法,极大地降低了计算复杂度,提高了模型性能。其次,在 Swin Transformer 的基础上加入全局的信息交互模块,加深了跨模态特征信息之间的表征能力,使模型能够获得更准确的图像分类准确率和更快的模型收敛速度。实验结果表明,该模型在公开数据集 ImageNet 上获得的分类准确率能达到 84.2%。本文方法相较于 Swin Transformer 图像分类方法,分类准确率提高了 2.8%。

关键词:图像分类;计算复杂度;信息交互;模型收敛
中图分类号:TP391 **文献标志码**:A

An Improved Swin Transformer Image Classification and Recognition Method

CHEN Cheng[1], GENG Xiaozhong[2], LIU Baijin[1], WANG Linen[1], HU Weixin[2]

(1. *School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin* 132022, *China*;
2. *School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun* 130012, *China*)

✉ 2468295244@qq.com; dq_gxz@ccit.edu.cn; 1692797200@qq.com; 3172876826@qq.com; 1051090429@qq.com

Abstract: This paper proposes an improved Swin Transformer image classification and recognition method to address the issue of excessive computational complexity in processing image tasks with the Transformer model. Firstly, Swin Transformer uses a patched image feature map processing method, which greatly reduces computational complexity and improves model performance. Secondly, by adding a global information exchange module on the basis of Swin Transformer, the representation ability between cross modal feature information is deepened, and the model can achieve more accurate image classification accuracy and faster model convergence speed. The results of this experiment indicate that the classification accuracy achieved by the model on the public dataset ImageNet can reach 84.2%. Compared to the Swin Transformer image classification method, the improved method has improved classification accuracy by 2.8%.

Key words: image classification; computational complexity; information interaction; model convergence

0 引言(Introduction)

计算机视觉建模的主流算法之前一直采用的是卷积神经网络(Convolutional Neural Networks,CNN),因为它首次实现了原始图片经过网络直接输出分类类别的端到端训练[1-2],例如具有革命性的模型 AlexNet[3]和 ResNet34[4]。

近年来,受 Transformer[5]架构在自然语言处理(NLP)任务中成功应用的启发,大量基于 Transformer 的模型逐渐被用来处理计算机视觉任务[6]。但是,图像识别中对图像分辨率的要求是非常高的,计算复杂度的大幅提升成为 Transformer 应用在图像处理任务中需要克服的难题[7]。为了降低计算的复

收稿日期:2023-05-18

杂度，Swin Transformer 借鉴了 Vision Transformer(ViT)中的补丁(Patch)化设计，在 Vision Transformer 的基础上提出 Shifted Window 模块，增强了区域特征信息建模能力。Shifted Window 被证明对全部 MLP(多层感知机)架构都是有益的。

Swin Transformer 中的补丁(Patch)化设计能够极大地降低计算复杂度，但过多归纳偏执可能会导致模态之间信息交互不充分。所以，本文加入 Large Kernel Block 模块进行全局的信息交流，根据有效感受野(Effective Receptive Field，ERF)理论，ERF 的大小与 Kernel 的大小和模型深度的平方根均成正比关系，因此该模块能够增强 Swin Transformer 的全局信息感知能力。本文在平衡模型复杂度和模型特征表述能力后提出的 Large Kernel Block 模块能够提升 Swin Transformer 的分类准确率。

1 算法原理(Algorithm principle)

1.1 自注意力

Self-Attention 的输入是特征映射 $\boldsymbol{x}\in\mathbb{R}^{C_{\text{in}}\times H\times W}$，其中 H 为高度、W 为权重、C_{in} 是通道数，自注意层的输出 $\boldsymbol{y}\in\mathbb{R}^{C_{\text{out}}\times H\times W}$ 由输入 $\boldsymbol{x}$ 经过以下方程映射得出：

$$y_{ij}=\sum_{h=1}^{H}\sum_{w=1}^{W}softmax(q_{ij}^{T}k_{hw})v_{hw} \tag{1}$$

其中：$\boldsymbol{q}=\boldsymbol{W}_Q\boldsymbol{x}$，$\boldsymbol{k}=\boldsymbol{W}_K\boldsymbol{x}$，$\boldsymbol{v}=\boldsymbol{W}_V\boldsymbol{x}$ 都是根据输入 $\boldsymbol{x}$ 计算得到的映射。q_{ij}、k_{ij}、v_{ij} 的 $i\in\{1,\cdots,H\}$，$j\in\{1,\cdots,W\}$。映射矩阵 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V\in\mathbb{R}^{C_{\text{in}}\times C_{\text{out}}}$ 是可学习的。如公式(1)所示，y_{ij} 是使用 $softmax$ 计算全局映射获得的。

1.2 非重叠窗口中的自注意力

标准的 Transformer 架构计算 Token 与其他所有 Token 之间的关系，全局的信息交互导致二次计算，提升了计算的复杂度，因此它在密集预测或者高分辨率的视觉任务中并不适用。Swin Transformer 模型将窗口划分为数个 $M\times M$ 大小的补丁(Patch)，全局 MSA 模块的计算复杂度如公式(2)所示，基于补丁(Patch)的 W-MSA 的计算复杂度如公式(3)所示：

$$\Omega(\text{MSA})=4hwC^2+2(hw)^2C \tag{2}$$

$$\Omega(\text{W-MSA})=4hwC^2+2M^2hwC \tag{3}$$

从以上公式可以明显地看到：公式(3)在处理高分辨率图片时，可以大幅度地降低计算的复杂度。

1.3 相对位置偏置

在计算自注意时，通常通过添加相对位置偏置 $\boldsymbol{B}\in\mathbb{R}^{M^2\times d}$ 来提高模型中相对位置的作用，由输入特征映射 $\boldsymbol{x}$ 得到输出 $\boldsymbol{y}$ 的方程如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}+\boldsymbol{B})\boldsymbol{V} \tag{4}$$

其中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}\in\mathbb{R}^{M^2\times d}$ 分别为 Attention 机制中的 $query$、key 和 $value$，d 是 $query/key$ 的维度，M^2 是特征图中被分成的补丁(Patch)数量，本文添加了可训练的相对位置偏置。经过实验发现，相对位置偏置能够增强模型的空间信息构造能力，提高模型性能。

2 网络模型分析(Network model analysis)

2.1 Swin Transformer 网络模型分析

Swin Transformer 的网络结构图如图 1 所示，它通过图片的分割模块 Patch Partition 将 $H\times W$ 的 RGB 图像输入后分割成独立的补丁(Patch)，每一个补丁(Patch)的特征被设置为原始像素 RGB 值的串联[8]。Patch Embeding 将原始的特征值映射到任意的维度中，最终将图像的大小处理为 $\frac{H}{4}\times\frac{W}{4}\times C$。

随着网络的深入，特征图通过 Patch Merging 层减少补丁(Patch)的数量。每次经过 Patch Merging 层图像的 H 与 W 都减少为之前的一半，特征通道扩大为 2 倍。网络中 stage 2、stage 3 和 stage 4 结构相同，目的是进行特征图的下采样，模型最后的输出像素块的大小为 $\frac{H}{32}\times\frac{W}{32}\times 8C$。

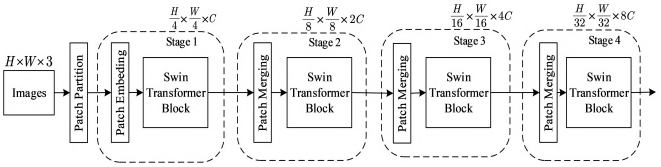

图 1 Swin Transformer 网络结构图

Fig. 1 Swin Transformer network structure diagram

Swin Transformer Block 网络结构图如图 2 所示，结构中包含在窗口中的多头注意力模块(W-MSA)与基于 Shifted Window 的多头自注意力模块(SW-MSA)。对两个模块进行层归一化(LayerNorm)，可以减少内部协变量偏移问题，最后经过多层感知机(MLP)[9]输出。

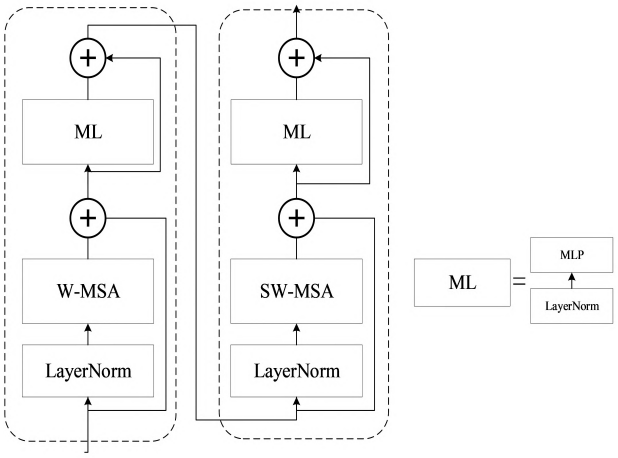

图 2 Swin Transformer Block 网络结构图

Fig. 2 Swin Transformer Block network structure diagram

对于密集像素的自注意力计算，其计算复杂度非常高，将其分割为小块后分别计算自注意力，能够大幅度地降低计算复杂度，Shifted Window 能够使不同补丁(Patch)之间进行信息交流，增强特征的全局感知。基于 Shifted Window 方法的连续 Swin Transformer 块计算公式如下：

$$\hat{z}^{l}=W\text{-}MSA(LN(z^{l-1}))+z^{l-1} \tag{5}$$

$$z^{l}=MLP(LN(\hat{z}^{l}))+\hat{z}^{l} \tag{6}$$

$$\hat{z}^{l+1}=SW\text{-}MSA(LN(z^{l}))+z^{l} \tag{7}$$

$$z^{l+1}=MLP(LN(\hat{z}^{l+1}))+\hat{z}^{l+1} \tag{8}$$

其中：$\hat{z}^{l}$ 和 z^{l} 为 Swin Transformer Block 中经过(S)W-MSA 模块和 MLP 模块的输出表征，W-MSA 和 SW-MSA 分别指的是不使用 Shifted Window 和使用 Shifted Window 的基于补丁(Patch)的多头自注意力。与 Shifted Window 一起使用的还有

掩码机制,其作用是在不使用额外的计算资源的情况下,实现局部的自注意力机制。Swin Transformer 模型将移位之后的补丁(Patch)按块标号,并对补丁(Patch)之外的信息进行屏蔽,从而实现了局部的自注意力。这样可以有效地节省计算资源,提升模型的运行效率。

2.2　改进的 Swin Transformer 网络模型分析

本文对 Swin Transformer 的结构进行优化。Swin Transformer 作为一个以补丁(Patch)注意力为基础的网络结构,加入 Shifted Window 节省了计算资源,但使用过多的归纳偏置会影响模型模态之间的表述能力,因此本文增加了 Swin Transformer 在全局上的信息交流[10]。

改进后的 Swin Transformer 网络结构图如图 3 所示,图像处理完之后,经过 Large Kernel Block 进行全局信息交流,然后进入 Swin Transformer Block 进行补丁(Patch)之间的信息交换。Large Kernel Block 分别添加在 stage 1、stage 2 和 stage 3 的 Swin Transformer Block 之前,能在网络对图片特征进行补丁(patch)化自注意力前增强模型中特征图的表述能力。

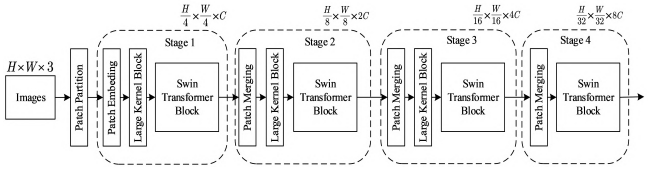

图 3　改进后的 Swin Transformer 网络结构图

Fig. 3 Improved Swin Transformer network structure diagram

本文在模块中加入的 Large Kernel Block 的网络结构图如图 4 所示,整体采用 Inverted Bottleneck 模块,按照中间宽两头小的信息通道进行设置。这样设计的目的是使处于高纬度的信息通过激活函数后,丢失的信息量会减少,可以保存更多的全局信息。

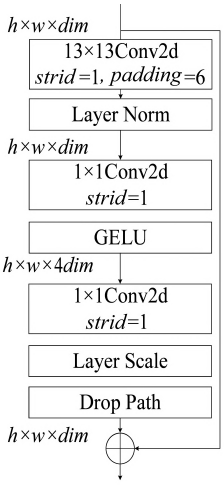

图 4　Large Kernel Block 网络结构图

Fig. 4 Large Kernel Block network structure diagram

模型中的激活函数没有使用卷积神经网络中最常用的 RELU,而是使用 Swin Transformer 中的 GELU。本文使用了更少的正则化(Normalization),只保留了深度卷积(Depthwise Convolution)后的 Normalization 层,并且本文使用的不是在卷积神经网络中常规的 Batch Normalization(BN),而是 Transformer 中经常使用的 Layer Normalization(LN)。经过测试,在模型 stage1、stage2 和 stage3 中,Large Kernel Block 的循环次数设置为[5,3,2]是最佳的。

Large Kernel Block 可以小幅扩大感受野(RF),提高模型对输出特征的感知能力,例如经典的大核卷积模型 RepLKNet,它的卷积核从[3,3,3,3]提升到[31,29,27,13],并在 ImageNet 与 ADE20K 数据集中表现出色。此外,Large Kernel Block 会给网络带来更多的形状偏执。大核卷积在小尺寸图片上的性能表现同样很好。RepLKNet 论文中的实验证明,将 MobileNetV2 最后一阶段的深度卷积尺寸提升到 7 和 13,应用在 ImageNet 数据集中进行验证,准确率分别提升了 0.24%和 0.21%,在 Cityscapes 数据集中,平均交并比(mIoU)的准确率分别提升了 1.9%和 2.31%。本文引入了大核卷积的设计,在进行 Shifted Window 自注意力之前进行多轮的大核 13×13 的卷积,增大了 Swin Transformer 中基于补丁(Patch)自注意力的感受野。

该模块在 Swin Transformer 的框架基础上加入了多层大核卷积。该模块作为特征图信息之间的交流层,在不增加计算复杂度的同时,提升了模型的收敛速度和分类准确率。

3　实验结果分析(Experimental results analysis)

3.1　数据集

该模型在 ImageNet 数据集中进行了多次小规模的实验,本文的数据集为 5 组从 ImageNet 数据集中随机抽取的 50 种类别的图片,测试集共 50 000 张图片,验证集为 15 000 张图片,图像的分辨率是 224×224。

3.2　实验细节

本文使用的是 PyTorch 深度学习框架,Python3.8 版本,NVIDIA RTX3080 显卡,10 GB 显存[11]。考虑到 Transformer 模型在小规模数据集中的表现差于大规模数据集的表现,本文进行 400 个 *epoch*,使用余弦退火学习率 CosineAnnealingLR,初始学习率设置为 0.000 005,AdamW 优化器。每批数据量(Batch Size)设置为 16,权重衰减(Weight Decay)设置为 0.05。本文没有设置 Mixup 数据增强,训练集和验证集的损失函数都是交叉熵损失 CrossEntropyLoss。

本文在相同的环境下进行实验,固定随机种子,确保对比实验中随机产生的数值相同。从图 5 可以看出,Swin-T 模型和改进后模型的 Loss 曲线都呈现相似的平滑下降趋势。改进后的 Swin-T 模型与 Swin-T 模型的 Loss 曲线的下降幅度大致相同,但最终保持在较低的水平,改进后的 Swin-T 模型能够从图片中学习到更多的特征图信息,模型的预测值更接近真实值,模型的性能更好。从图 6 可以看出,改进后的 Swin-T 模型的分类准确率达到 84.2%相较于 Swin-T 的 81.4%有稳步提升。模型损失对比图(图 5)和模型准确率对比图(图 6)可以看出,改进后的 Swin-T 模型能够更快、更准确地学习到有用的特征信息,是因为 Large Kernel Block 的加入能够提升模型的感受野。

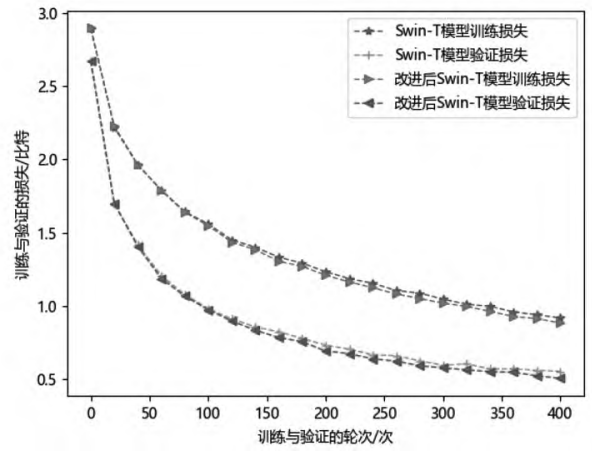

图 5 模型损失对比图

Fig. 5 Comparison of model losses

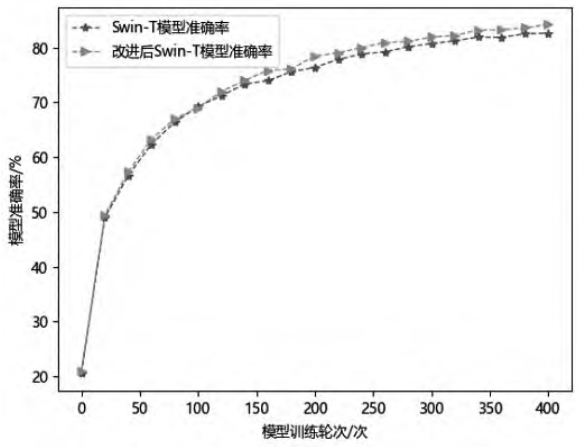

图 6 模型准确率对比图

Fig. 6 Comparison of model accuracy

表 1 是不同主干网络模型的参数对比，从中可以看出改进后的 Swin-T 在准确率上相较于常用 Transformer 模型略有增加。

表 1 不同主干网络模型的参数对比

Tab.1 Comparison of parameters of different backbone networks

模型	ImageNet 尺寸	参数/MB	浮点数/GB	ImageNet Top-1 ACC
DeiT-S	224×224	22	4.6	79.8
Swin-T	224×224	29	4.5	81.4
改进后 Swin-T	224×224	34	5.2	84.2

4 结论(Conclusion)

本文在 Swin Transformer 的基础上加入了全局特征交流模块 Large Kernel Block，并对 Kernel 层做出一些适合 Transformer 的改动。Large Kernel Block 增强了模型在特征提取阶段对于全局信息的感知，并且它的计算效率高于通常使用的卷积模块的计算效率，存算比也更高。实验证明这种设计能够显著提升模型在小规模数据集上的收敛速度，并且能够取得更好的分类效果。

基于 Transformer 的特征表示框架强大的跨模态表征能力引起了包括计算机视觉和自然语言处理等多个人工智能子领域的广泛关注，并被认为是目前实现通用智能的最佳框架。然而，将其应用在视觉领域还有很多问题有待进一步研究，主要是平衡跨模态表征能力和视觉领域的归纳偏置。在不使用过多归纳偏置影响模型模态之间的表述能力的前提下，使计算量减少到可接受范围，将会是后续 Transformer 在视觉领域的研究热点。

参考文献(References)

[1] 池亚平，岳梓岩，林雨衡. 基于 Transformer 的 SM4 算法工作模式识别[J]. 计算机工程，2023，49(9)：109-117.

[2] 邵闻睿，汪远，张羽菲，等. 基于改进 DenseNet 的图像分类[J]. 中国宽带，2022，18(8)：64-66.

[3] 黄清，方木云. 一种基于 HMM 算法改进的语音识别系统[J]. 重庆工商大学学报(自然科学版)，2022，39(5)：56-61.

[4] 张莉，丁毛毛，李玮，等. 基于决策树算法的客服终端冗余数据迭代消除方法[J]. 计算技术与自动化，2022，41(4)：118-122.

[5] HASSAN M M，HASSAN M R，HUDA S，et al. A predictive intelligence approach to classify brain-computer interface based eye state for smart living[J]. Applied soft computing，2021，108：107453.

[6] SOSULSKI J，TANGERMANN M. Introducing block-Toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain-computer interfaces[J]. Journal of neural engineering，2022，19(6)：166-176.

[7] 李远，时旭，杨正春，等. 面向高光谱医学图像分类的空-谱自注意力 Transformer[J]. 光学精密工程，2023，31(18)：2752-2764.

[8] PARK Y，CHUNG W. Frequency-optimized local region common spatial pattern approach for motor imagery classification[J]. IEEE transactions on neural systems and rehabilitation engineering，2019，27(7)：1378-1388.

[9] 李映松，杨爱英，刘轩，等. 基于 Transformer 改进的 Faster-Rcnn 仓储箱体检测算法[J]. 自动化与仪器仪表，2022(8)：1-6.

[10] 赵锟，余添，周立俭，等. 基于 CNN-Transformer 的街景图像分类[J]. 青岛理工大学学报，2023，44(3)：146-152.

[11] YI S，LIU X，LI L，et al. Infrared and visible image fusion based on blur suppression generative adversarial network[J]. Chinese journal of electronics，2023，32(1)：177-188.

作者简介：

陈 成(2000-)，男，硕士生。研究领域：脑机接口技术及应用，图像识别。

耿晓中(1972-)，女，博士，教授。研究领域：脑机接口技术及应用，云计算。本文通信作者。

刘柏进(1999-)，男，硕士生。研究领域：时间序列预测。

汪林恩(1997-)，男，硕士生。研究领域：脑机接口技术及应用，嵌入式开发，深度学习技术。

户唯新(1999-)，男，硕士生。研究领域：脑机接口技术及应用，图像识别。