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 ,SwinTransformer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 SwinTransformer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 SwinTransformer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 ImprovedSwin TransformerImageClassification andRecognition Method

CHEN Cheng1 , GENG Xiaozhong2 , LIU Baijin1 , WANG Linen1 , HU Weixin2

(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% .

Keywords: 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(ConvolutionalNeuralNetworks,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 ,SwinTransformer借鉴了 VisionTransformer(ViT) 中的 补丁(Patch) 化 设 计 , 在 Vision Transformer的 基 础 上 提 出 ShiftedWindow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。

SwinTransformer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 LargeKernelBlock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 LargeKernelBlock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

C ×H×W

Self-Attention的输入是特征映射 x∈ℝ in ,其中 H 为高 度、W 为 权 重、Cin 是 通 道 数 , 自 注 意 层 的 输 出 y∈

C ×H×W

ℝ out 由输入 x经过以下方程映射得出 :

yij = softmax(qkhw )vhw (1) q、,,{输}

C ×C

WQ、WK、WV ∈ℝ in out 是可学习的 。如公式(1)所示 ,yij 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 。SwinTransformer模型将窗口划分为数个 M ×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 :

ΩW(M-C 

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

2

在计算自注意时 ,通常通过添加相对位置偏置 B∈ℝM ×d 来提高模型中相对位置的作用 ,由输入特征映射 x得到输出y 的方程如下 :

A**t**ention(,K,V)=softmax(QKT/ d+B)V (4)

其中 :Q、K、V∈ℝM ×d 分别为 Attention机制中的 query、key 和 value,d是query/key的维度 ,M2 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(Networkmodelanalysis) 2. 1Swin Transformer网络模型分析

SwinTransformer的网络结构图如图 1所示 ,它通过图片 的分割模块 PatchPartition将 H×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] 。PatchEmbeding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× ×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 stage2、 stage3和 stage4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××8C。

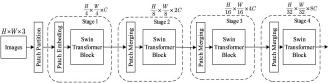


图 1 SwinTransformer网络结构图

Fig. 1SwinTransformernetwork structurediagram

SwinTransformerBlock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。

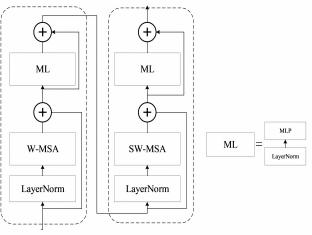


图 2 SwinTransformerBlock网络结构图

Fig. 2SwinTransformerBlock network structurediagram

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 ,ShiftedWindow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 ShiftedWindow方法的连续

SwinTransformer块计算公式如下 :

W=l=-SMMWSAL-MP( 

其中 :l 和 zl l+过(S)W-

模块和 MLP模块的输出表征 ,W-MSA和 SW-MSA分别指的 是不使用 ShiftedWindow和使用 Shifted Window 的基于补丁 (Patch)的多头自注意力 。与 ShiftedWindow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 SwinTransformer 的结构进行优化。SwinTransformer 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 SwinTransformer网络结构图如图 3所示 ,图像 处理完之后 ,经过 LargeKernelBlock进行全局信息交流 ,然后 进入 SwinTransformerBlock进行补丁(Patch) 之间的信息交 换 。LargeKernelBlock分别添加在 stage1、stage2和 stage3 的 SwinTransformerBlock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。

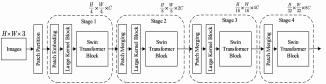


图 3 改进后的 SwinTransformer网络结构图

Fig. 3ImprovedSwinTransformernetwork structurediagram

本文在模块中加入的 LargeKernelBlock的网络结构图如 图 4所示 ,整体采用InvertedBottleneck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。

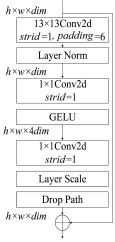


图 4 LargeKernelBlock网络结构图

Fig. 4LargeKernelBlock network structurediagram

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 SwinTransformer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 BatchNormalization(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 KernelBlock的循环次数设 置为[5,3,2]是最佳的。

LargeKernelBlock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 KernelBlock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 ShiftedWindow 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 SwinTransformer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(Experimentalresultsanalysis)

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 50000张图片 ,验证集为 15000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 ,10GB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. 000005, AdamW 优化器 。每批数据量 (BatchSize)设置为 16,权重衰减(WeightDecay)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 LargeKernelBlock的加入能够提升模型的感 受野。

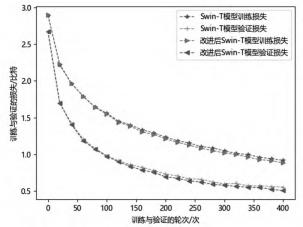


图 5 模型损失对比图

Fig. 5Comparison ofmodellosses

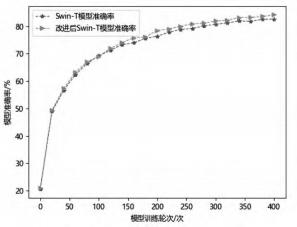


图 6 模型准确率对比图

Fig. 6Comparison ofmodelaccuracy

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-1ACC |
| 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 SwinTransformer的基础上加入了全局特征交流模 块 LargeKernelBlock,并对 Kernel层做出一些适合 Transformer 的改动 。LargeKernelBlock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,HUDAS,etal. A predic- tiveintelligence approach to classify brain-computer inter- facebased eyestateforsmartliving[J]. Applied softcom- puting,2021,108:107453.

[6] SOSULSKI J, TANGERMANN M. Introducing block-To- eplitz covariancematricesto remasterlineardiscriminantanal- ysisfor event-related potentialbrain-computer interfaces[J]. Journalofneuralengineering,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 spatialpattern approachformotorimageryclassi- fication[J]. IEEE transactionsonneuralsystemsand reha- bilitation engineering,2019,27(7) :1378-1388.

[9] 李映松 ,杨爱英 , 刘轩 ,等. 基于 Transformer改进的 Fas- ter-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] YIS, LIU X, LIL, etal. Infrared and visible image fusion basedonblursuppressiongenerativeadversarialnetwork[J].

Chinesejournalofelectronics,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。