1

1.1 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

1.1.1 问题描述

1. 定义

Swin Transformer是一种新型的视觉Transformer模型，通过分层次的移动窗口方法来计算表示，从而应对视觉领域的挑战，如视觉实体的尺度变化大和图像像素的高分辨率。该模型在计算复杂度上与图像尺寸呈线性关系，并适用于多种视觉任务。

2. 分类

根据视觉任务的不同，Swin Transformer可以应用于图像分类、目标检测和语义分割任务。在这些任务中，它通过构建分层特征图和引入跨窗口连接，实现了高效的计算和优异的性能。

3. 评价指标

在图像分类任务中，使用准确率（Accuracy）作为评价指标。在目标检测任务中，主要使用box AP和mask AP作为评价指标。在语义分割任务中，使用mIoU（平均交并比）作为评价指标。

4. 常用的数据集

论文中使用了ImageNet-1K、COCO和ADE20K数据集进行实验和评估。这些数据集分别用于图像分类、目标检测和语义分割任务。

5. 发展历史

Swin Transformer的发展背景主要是受到NLP领域中Transformer成功的启发。随着视觉领域中Transformer模型的不断探索，研究者们尝试解决视觉任务中特有的挑战，如高分辨率图像的处理复杂度和视觉实体的尺度变化问题。

1.1.2 论文摘要

本文提出了一种名为Swin Transformer的新型视觉Transformer，通过分层次的移动窗口方法计算自注意力表示。该模型在计算复杂度上与图像大小呈线性关系，并在多种视觉任务中表现出色，显著优于现有方法。

1.1.3 论文动机

在过去几年中，卷积神经网络（CNN）作为计算机视觉领域最重要的技术之一，在图像分类等众多任务中取得了良好的表现。然而，随着任务的复杂性增加，早期的卷积神经网络模型（如LeNet和AlexNet）相对较浅，难以处理更复杂的图像特征，仍然存在过拟合、训练时间长等问题。此外，Transformer在自然语言处理领域的成功引发了研究者对其在视觉领域应用的兴趣，但视觉任务中的高分辨率和大尺度变化使得直接应用Transformer面临挑战。

因此，这篇文章的目标和动机一方面是探索视觉Transformer在大规模图像任务中的表现，探索分层次的移动窗口方法对于视觉Transformer的影响；另一方面是提出一种新的、更有效的分层次视觉Transformer结构，称为Swin Transformer，通过移动窗口方法来提高计算效率，解决以往模型在处理高分辨率图像时的计算复杂度问题，并在多种视觉任务上进行测试和评估。

本文通过对已有的Transformer模型进行改进和优化，构建分层次的移动窗口机制，并在ImageNet、COCO和ADE20K数据集上进行训练和测试，比较不同结构的性能差异。作者认为，分层次的移动窗口方法可以提高模型对图像特征的抽象能力和表示能力，更好地捕捉图像中的细节和全局信息，从而提高分类、检测和分割任务的准确率。同时，本文还探讨了在不同分辨率和计算复杂度下，移动窗口方法对视觉Transformer模型性能的影响。

1.1.4 实验方法

本文提出的Swin Transformer的实验方法主要包括以下几个部分：分块模块、Swin Transformer块的设计、移动窗口自注意力机制以及相对位置偏置的应用。

1. 分块模块

输入图像首先被分割成不重叠的4×4小块（token），每个小块作为一个独立的输入单元，这样能够有效减少计算复杂度。线性嵌入层将这些小块的原始像素RGB值投影到特定维度C，从而进一步降低计算开销。

2. Swin Transformer块的设计

Swin Transformer块由移动窗口自注意力模块和多层感知机（MLP）模块组成。每个移动窗口自注意力模块包括LayerNorm层、自注意力计算和残差连接，而MLP模块则由两层全连接层和一个GELU激活函数组成。这种设计能够增强模型的稳定性和训练效果。移动窗口自注意力模块在每个窗口内计算自注意力，并通过移动窗口划分引入跨窗口连接，提高模型的表达能力。

3. 移动窗口自注意力机制

在传统的Transformer架构中，自注意力机制通常在全局范围内计算，导致计算复杂度随输入大小呈二次增长。为了解决这个问题，Swin Transformer在局部窗口内计算自注意力，每个窗口均匀地分割图像且不重叠。为了引入跨窗口连接，在连续的自注意力层之间采用移动窗口划分策略，使得新窗口跨越前一层的边界，从而在窗口之间引入连接。这种方法不仅提高了计算效率，还增强了模型的全局表示能力。

4. 相对位置偏置的应用

在计算自注意力时，引入相对位置偏置，可以帮助模型捕捉到局部窗口内的位置信息，从而进一步提高自注意力机制的有效性。

5. 实验设置

本文在ImageNet-1K、COCO和ADE20K数据集上进行了广泛的实验。在训练过程中，使用了随

机梯度下降（SGD）优化器，并采用了不同的数据增强技术来扩充训练集。此外，为了减少过拟合，还使用了dropout技术。

6. 分类任务

在ImageNet-1K数据集上，模型进行了300个epoch的训练，使用AdamW优化器，初始学习率设置为0.001，权重衰减为0.05，批量大小为1024。在此基础上，进一步进行了ImageNet-22K预训练和微调。

7. 检测和分割任务

在COCO数据集上，使用了Mask R-CNN框架进行目标检测和实例分割，训练过程中采用了3x schedule（36个epoch）。在ADE20K数据集上，使用UperNet框架进行语义分割训练。

1.1.5 实验结果

在多个视觉任务上的实验结果显示，Swin Transformer在图像分类、目标检测和语义分割任务中表现出色，显著优于现有的最佳方法。具体结果如下：

1. 图像分类任务

在ImageNet-1K数据集上，Swin Transformer取得了87.3%的Top-1准确率，超过了许多现有的模型。例如，与其他流行模型相比，Swin Transformer不仅在准确率上更高，而且在计算效率上也表现出色。

图示：Swin Transformer与其他流行模型在ImageNet-1K数据集上的分类性能比较。

2. 目标检测任务

在COCO数据集上，Swin Transformer在box AP和mask AP上分别达到了58.7和51.1。这些结果显示，Swin Transformer在目标检测任务中不仅准确度高，而且推理速度快，超越了许多现有的检测模型。

图示：Swin Transformer与其他流行模型在COCO数据集上的目标检测性能比较。

3. 语义分割任务

在ADE20K数据集上，Swin Transformer达到了53.5的mIoU，比之前的最佳方法提高了显著的数值。这表明，Swin Transformer在捕捉图像细节和处理复杂场景方面具有很强的能力。

图示：Swin Transformer与其他流行模型在ADE20K数据集上的语义分割性能比较。

4. 模型的宽度和分辨率调整

作者还评估了Swin Transformer在不同宽度和分辨率设置下的性能。在降低宽度和分辨率的情况下，模型的计算复杂度和延迟显著减少，而准确率保持在一个较高水平。例如，在降低宽度乘数和分辨率乘数的情况下，Swin Transformer仍然能够保持较高的准确率，同时大大减少了模型大小和推理时间。

图示：不同宽度和分辨率设置下的Swin Transformer性能比较。

5. 跨平台性能评估

Swin Transformer在不同硬件平台上也表现出色。例如，在CPU、GPU和FPGA等平台上，Swin Transformer在推理时间和准确率方面都表现优异。具体来说，在CPU上运行时，Swin Transformer的推理时间显著短于许多现有模型，而在GPU上，其推理时间进一步减少，同时保持高准确率。
图示：Swin Transformer在不同硬件平台上的性能比较。

通过这些实验结果可以看出，Swin Transformer在多个视觉任务中均表现出了优异的性能和高效的计算能力，证明了其作为通用视觉Transformer骨干网络的巨大潜力。

1.1.6 实验总结

本实验主要研究了视觉Transformer在多个视觉任务中的表现，提出了一种新的分层次移动窗口自注意力机制来优化模型性能。通过在ImageNet-1K、COCO和ADE20K数据集上进行实验，论文作者证明了Swin Transformer相对于传统的视觉Transformer模型具有更高的性能和更高的计算效率。

具体来说，本实验采用了严谨、全面、系统化的实验设计方法，包括对不同分辨率和窗口配置的网络结构进行训练和测试、对比实验验证移动窗口自注意力机制相对于传统全局自注意力机制的优越性、在多个计算机视觉任务中测试Swin Transformer的性能表现等。通过这些实验，论文作者得出了以下结论：

•分层次移动窗口自注意力机制显著提高了模型的计算效率：相对于传统全局自注意力机制，分层次移动窗口自注意力机制能够在保持高性能的同时显著降低计算复杂度，使得模型在处理高分辨率图像时更加高效。

•随着网络深度增加，使用分层次移动窗口自注意力机制可以显著提高模型性能：实验结果表明，分层次设计和移动窗口机制使得模型能够更好地捕捉不同尺度的图像特征，提高了分类、检测和分割任务的准确率。

•SwinTransformer在多个视觉任务中取得了优异成绩：在ImageNet-1K数据集上的图像分类任务、COCO数据集上的目标检测任务以及ADE20K数据集上的语义分割任务中，Swin Transformer均表现出色，显著超过了现有的最佳方法。

总之，本实验为解决视觉Transformer在高分辨率图像处理中的计算复杂度问题提供了一种新的思路和方法，对于推动计算机视觉领域的发展具有重要意义。未来的研究可以进一步探索Swin Transformer在其他视觉任务中的应用，并继续优化其结构和性能。