**Title:** Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

The code and models are publicly available at https://github.com/microsoft/Swin-Transformer.

**Abstract:**

This paper presents a new vision Transformer, called Swin Transformer, that capably serves as a general-purpose backbone for computer vision.

They propose a hierarchical Transformer whose representation is computed with Shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection

Its performance surpasses the previous state-of-the art by a large margin of +2.7 box AP and +2.6 mask AP on COCO, and +3.2 mIoU on ADE20K.

这篇论文提出了新的ViT，叫Swin Transformer，比之前的表现都好，可以作为CV模型的骨干

它用Shifted windows的方法，将自注意力计算限制在了不重叠的本地窗口上，同时允许跨窗口通信，这给予了它更好的效率

**Intro:**

BACKGROUND:

Modeling in computer vision has long been dominated by convolutional neural networks (CNNs).

On the other hand, the evolution of network architectures in natural language processing (NLP) has taken a different path, where the prevalent architecture today is instead the Transformer. Its tremendous success in the language domain has led researchers to investigate its adaptation to computer vision.

CNN占据了CV主流，NLP走的路不同，他们用transformers，研究人员希望将transformer用于CV领域

PROBLEM:

Scale: Word tokens that serve as the basic elements of processing in language but unsuitable for vision applications.

Compute:Another difference is the much higher resolution of pixels in images compared to words in passages of text. This would be intractable for Transformer on high-resolution images, as the computational complexity of its self-attention is quadratic to image size.

规模：现有的Transformer中token都是固定长度的，无法处理视觉问题。

计算：图像分辨率比文本段落中的单词相比高很多，而且注意力计算复杂度是图像大小的二次方，对于高分辨率图像Transformer是不可处理的。

APPROACH:

Swin Transformer constructs a hierarchical representation by starting from small-sized patches (outlined in gray) and gradually merging neighboring patches in deeper Transformer layers. The linear computational complexity is achieved by computing self-attention locally within non-overlapping windows that partition an image (outlined in red). The number of patches in each window is fixed, and thus the complexity becomes linear to image size. The shifted windows bridge the windows of the preceding layer, providing connections among them that significantly enhance modeling power.

在深层Transformer层对每个小块进行合并构建层次表示，对图像大小复杂度是线性的是将图片进行分区，每个小块的数量在移位窗口都是固定的，每个窗口单独进行自注意力计算。移位的窗口桥接了前一层的窗口，连接了它们使得建模能力显著提高。

RESULT:

The proposed Swin Transformer achieves strong performance on the recognition tasks of image classification, object detection and semantic segmentation.

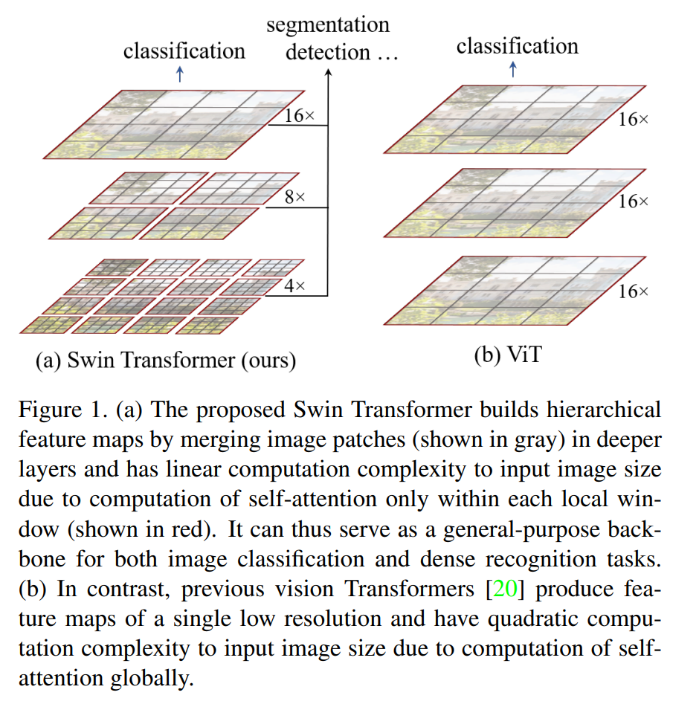
Swin Transformer在图像分类，目标检测和语义分割方面都有强劲性能。

**Conclusion:**

This paper presents Swin Transformer, a new vision Transformer which produces a hierarchical feature representation and has linear computational complexity with respect to input image size. Swin Transformer achieves the state-of-the-art performance on COCO object detection and ADE20K semantic segmentation, significantly surpassing previous best methods.

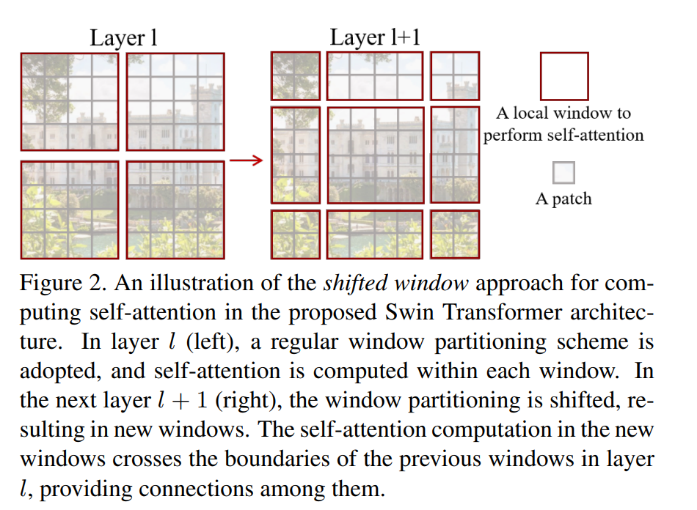
Swin Transformer 创建了一个层次特征显示并具有对图像大小有线性复杂度的Transformer。达到了目前最先进的水平。

**Figure:**

Swin Transformer通过从小的patch（以灰色轮廓表示）开始，并逐渐在更深层的Transformer层中合并相邻的patch来构建层次化的表示。

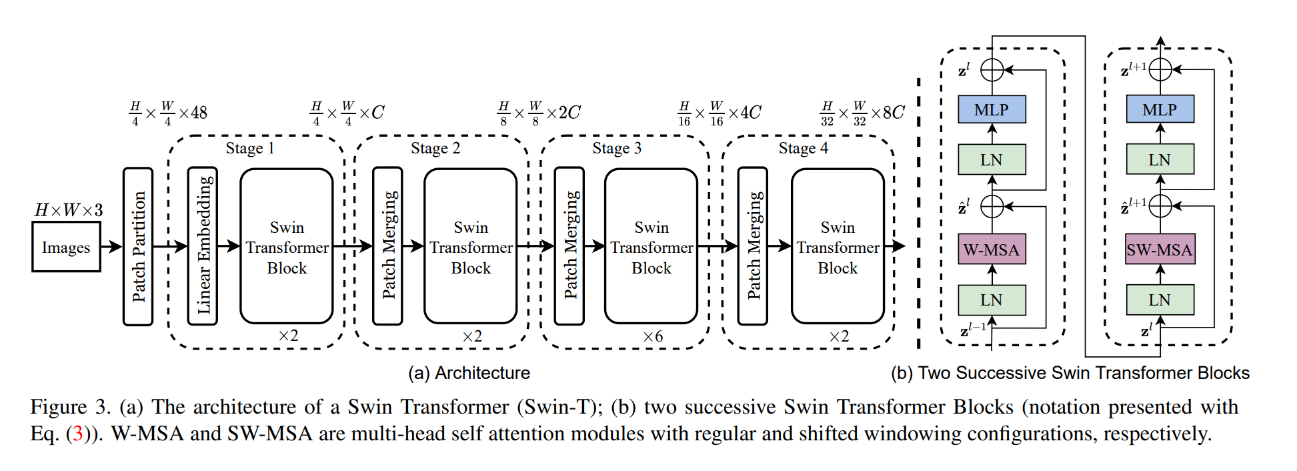
由于只在每个局部窗口内计算自注意力，因此其计算复杂度与输入图像的大小呈线性关系。

相比之下，之前的视觉Transformer[20]产生单个低分辨率的特征图，并且由于全局计算自注意力，其计算复杂度与输入图像的大小呈二次关系。



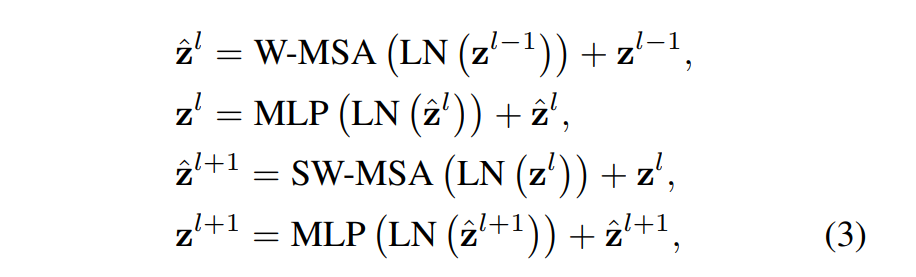
Swin Transformer采用Shift window方案，在l层采用常规分区方案，在l+1层窗口发生移动，新的窗口中计算的自注意力跨越层l中先前的窗口边界，从而在这些窗口之间建立了联系。

通过跨步采样降低H，W，导致通道数为4倍，然后再通过1\*1的卷积降低通道数为2倍



Swin Transformer（Swin-T）的结构；两个连续的Swin Transformer Block（使用式（3）进行表示：下图）。W-MSA和SW-MSA是具有常规和Shifted windowing配置的多头自注意力模块。

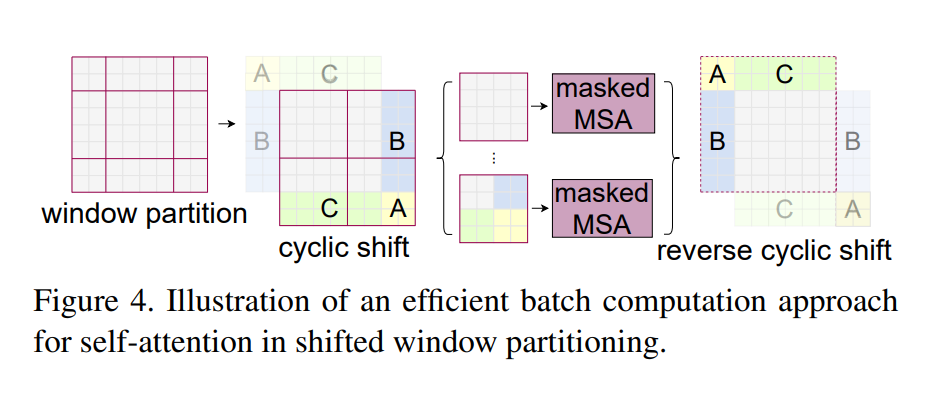
W-MSA和SW-MSA总是成对出现，因为要完成窗口间的注意力交互



残差连接

LayerNorm层，加速收敛

MLP全连接层

一种有效的计算Shift window分区自注意力的批处理方法的说明

由朝左上方向的cyclic-shifting完成

发生这种位移之后，批量窗口可由在特征图上彼此不相邻的若干子窗口构成

采用遮蔽机制将自注意力计算限制在每个子窗口内。采用循环位移之后，批量窗口的数量仍然与常规窗口分区的数量相同，因此也是高效的。

**Relate Work:**

**CNNs**

**Self-attention based backbone architectures:** some works employ self-attention layers to replace some or all of the spatial convolution layers in the popular ResNet. However, their costly memory access causes their actual latency to be significantly larger than that of the convolutional networks. Instead of using sliding windows, we propose to shift windows between consecutive layers, which allows for a more efficient implementation in general hardware.

一些工作将ResNet中卷积层替换成了自注意力层，但是有昂贵的存储器访问。Shift window允许在通用硬件中高效实现

**Self-attention/Transformers to complement CNNs**

**Transformer based vision backbones**

**Method:**

**Overall Architecture:** Figure 3

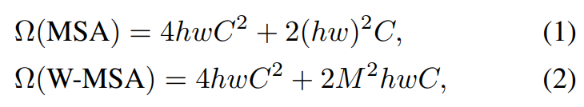
Swin Transformer block Swin Transformer is built by replacing the standard multi-head self attention (MSA) module in a Transformer block by a module based on shifted windows (described in Section 3.2), with other layers kept the same.

将MSA替换为基于shifted windows的module

**Shifted Window based Self-Attention:（关键）**

1.Self-attention in non-overlapped windows:

Supposing each window contains M × M patches, the computational complexity of a global MSA module and a window based one on an image of h × w patches are:

 MSA和W-MSA的计算复杂度 对比，可以看到MSA为hw的平方，W-MSA不是，所以全局注意力计算不适合hw过大的图像。

2.Shifted window partitioning in successive blocks：

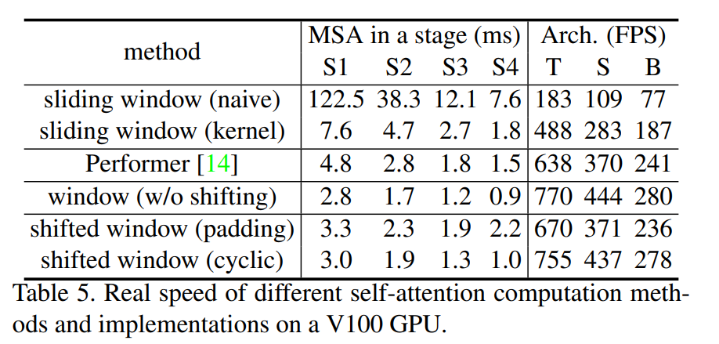
先将图片从左上角开始划分，8\*8的特征图划分成2\*2个窗口，每个窗口大小为4\*4（M=4），然后窗口移动（M/2，M/2）个像素。详细见Figure4

通过Swin Transformer block的计算，连接了相邻不重叠的窗口，它被发现在图像分类，目标检测和语义分割上是有效的。

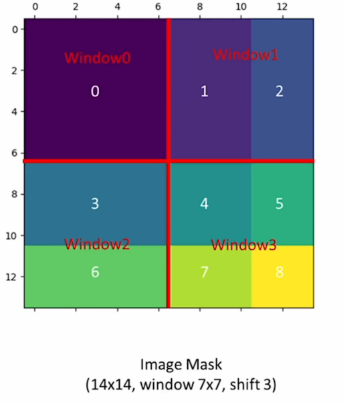
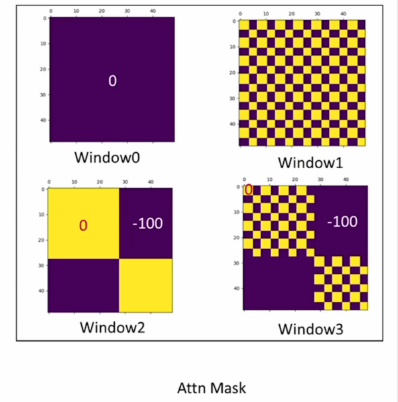
3. Efficient batch computation for shifted configuration：

因为窗口移位，若是简单的pad较小的窗口，那么就会使计算量增大，所以使用了循环移位的方法，见Figure4

这种方法的低延迟如下表所示：



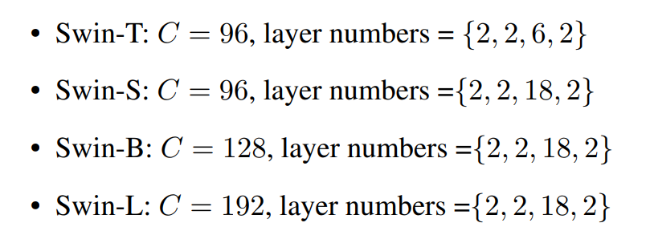
循环移位后，通过巧妙地Mask操作，降低了计算复杂度（矩阵乘法，将不应该计算注意力的值mask掉）



Relative position bias：

Swin-T 在计算 Attention 的时候做了一个相对位置编码

**Architecture Variants：**

****

**QUESTIONS：**

-what did author try to accomplish?

Transformer’s tremendous success in NLP has led researcher to investigate its adaption to computer version. Author tried to created a new transformer which can be used in CV domain.

-what the key elements of the approach?

According to Ablation Study in paper, the answer is The Shifted windows and Relative position bias and self-attention methods.

-what can you use yourself?

Firstly, I can use this model by this paper for study or development. Secondly, I learn much more knowledge about the self-attention, transformer models and the performance of Swin Transformers. Finally, this paper teach me how to introduce other domain’s model to relative domain and how to arrange my paper and so on.

-what other reference do you want to follow?

Mybe other models or how to use models to improve our life quality.