—— x ——

Swin Transformer

(Shifted Window Transformer)

Limitation of ViT:

1) Quadratic computational complexity,

* Attention complexity scales $O(N^2)$
* N = number of patches.
* For images, $N \propto H \times W$

→ Imagine image with 9 patches
→ Total attention scores will be 9×9

Goal of Swin Transformer:

* Make transformers a general purpose backbone for vision task while achieving,

→ linear computational complexity w.r.t img size

→ strong performance on classification, detection and segmentation.

* It restricts self attention to local windows.

Architecture:

I/p Image (H×W×3)

↓

Patch Partition (4×4) --→ Result : H/4 × W/4 resolution

↓ After flattening = $\frac{H}{4} \times \frac{W}{4} \times 48$

stage-1

Linear Embedding --→ Project to C, Result = H/4 × W/4 × C

↓

Swin Transformer Block ×2 --→ Result : H/4 × W/4 × C

↓

stage-2

Patch Merging --→ Result : H/8 × W/8 × 2C

↓

Swin Transformer Block ×2 --→ Result : H/8 × W/8 × 2C

↓

stage-3

Patch Merging --→ Result : H/16 × W/16 × 4C

↓

Swin Transformer Block ×6 --→ Result : H/16 × W/16 × 4C

↓

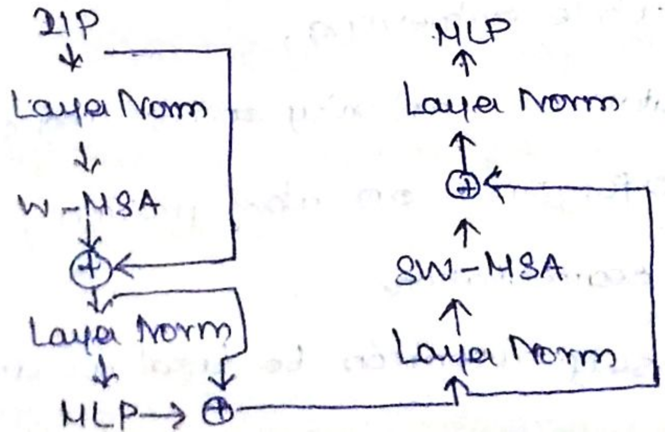
stage-4

Patch Merging --→ Result : H/32 × W/32 × 8C

↓

Swin Transformer Block ×2 --→ Result : H/32 × W/32 × 8C

Swin Transformer Block:



Patch Partition

Step 1: Image → Partition

* I/P image: $H \times W \times 3$
* Patch size: 4×4
* Each patch contains: $4 \times 4 \times 3 = 48$ values.

Step 2: Flattening:

* Each patch → 48 dim vector
* Resulting dim → $H/4 \times W/4 \times 48$

Linear Embedding:

* Its a linear projection converts $48 \to C$
* Resulting dim: $H/4 \times W/4 \times C$

Patch Merging:

* It reduces spatial size while increasing channels.

↳ Working:

→ Take 2×2 groups of neighboring patches

→ Concatenate to channels: $C + C + C + C = 4C$

→ Apply linear projection: $4C \to 2C$

* Result = Height ↓ 2, Width ↓ 2, Channels ↑ 2

Window-Based Self Attention (W-MSA):

* Instead of global attention, divide images into non-overlapping windows and compute self attention within each window only.

→ Let window size = $M \times M$

→ No of windows $\approx \dfrac{H \times W}{P^2 \times M^2}$

Total attention complexity $O\left(\frac{HW}{P^2} \cdot M^2\right)$

Shifted Window Attention (SW-MSA):

* Shifting the window in a cyclic manner let patches that move out of one side and re-enter from the opposite side.

* This enables cross window communication indirectly.

Attention Formula in Swin:

Regular Window Attention,

$$\text{Attention} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V$$

where,

$B \rightarrow$ relative position bias.

Shifted Window Attention,

$$\text{Attention} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B + M\right)V$$

where,

- $M = 0$ if the token are same window
- $M = -\infty$ otherwise.

Relative Position Bias.

* Unlike ViT where absolute position embedding was implemented, in swin transformer a learnable bias is added

* It is based upon relative displacement $(\Delta x, \Delta y)$ for a 7×7 window,

$\rightarrow \Delta x, \Delta y \in [-6, +6]$

$\rightarrow$ Total unique relative position : $13 \times 13 = 169$

* There is only 169 learnable parameters, instead of $49 \times 49 = 2401$

Classification in Swin Transformer: [Although Swin Transformer is General Purpose]

* There is no CLS Token, instead the final stage features are global average pooled and passed to classification head.

— x —