linear computational complexity to the image size.

SWIN TRANSFORMER:

↳ Shifted Window Transformer is a type of Vision Transformer designed to overcome some of the limitations of Vision Transformer.

↳ It uses window-based self attention, with a unique mechanism called shifted windows to improve efficiency and scalability,

↳ The main goal of Swin Transformer is to make self attention scalable for high resolution images without losing performance. —(a)

↳ It does (a) by partioning the image into non-overlapping windows and shifting windows b/w the layer, which reduces the computational cost compared to ViT, which computes attention globally across the entire image.

Important Note:-

(I) Instead of attending to every pixel globally (like in ViT), the image is divided into windows of local patches. and self attention is applied within each window, The windows shift b/w layers to ensure information is shared across patches.

↳ this approach make computational complexity $\propto$ image size

(II) Reduces computation → complexity

↳ ViT approach has a quadratic complexity.

SWIN TRANSFORMER ARCHITECTURE:-

(i) Patch embeddings (similar to ViT).

(ii) Hierarchial Transformer Blocks (unlike ViT where all patches are processed equally/simultaneously)

CONCEPT OF PATCH MERGING

The image that is divided into small patches (like ViT). These patches are fed into the transformer and processed using self attention within windows.

Multiple patches are merged to form larger patches, reducing the no. of patches but increasing the amount of information in each patch

The patches are merged again, further reducing resolution now, the patches are big images and the transformer makes the final prediction.

→ The hierarchial transformer blocks means the resolution of feature maps progressively

These transformer blocks however contain the same layer (coherent).

(i) Window Based Self Attention.

(ii) Shifted Window Self Attention.

(iii) MLP (Multilayer Perceptron) block. for further processing.

Step no.1 :

The images are split into a non-overlapping windows.

Selfattention is applied within each window.

No patch from one window can interact with another window.

Step no.2

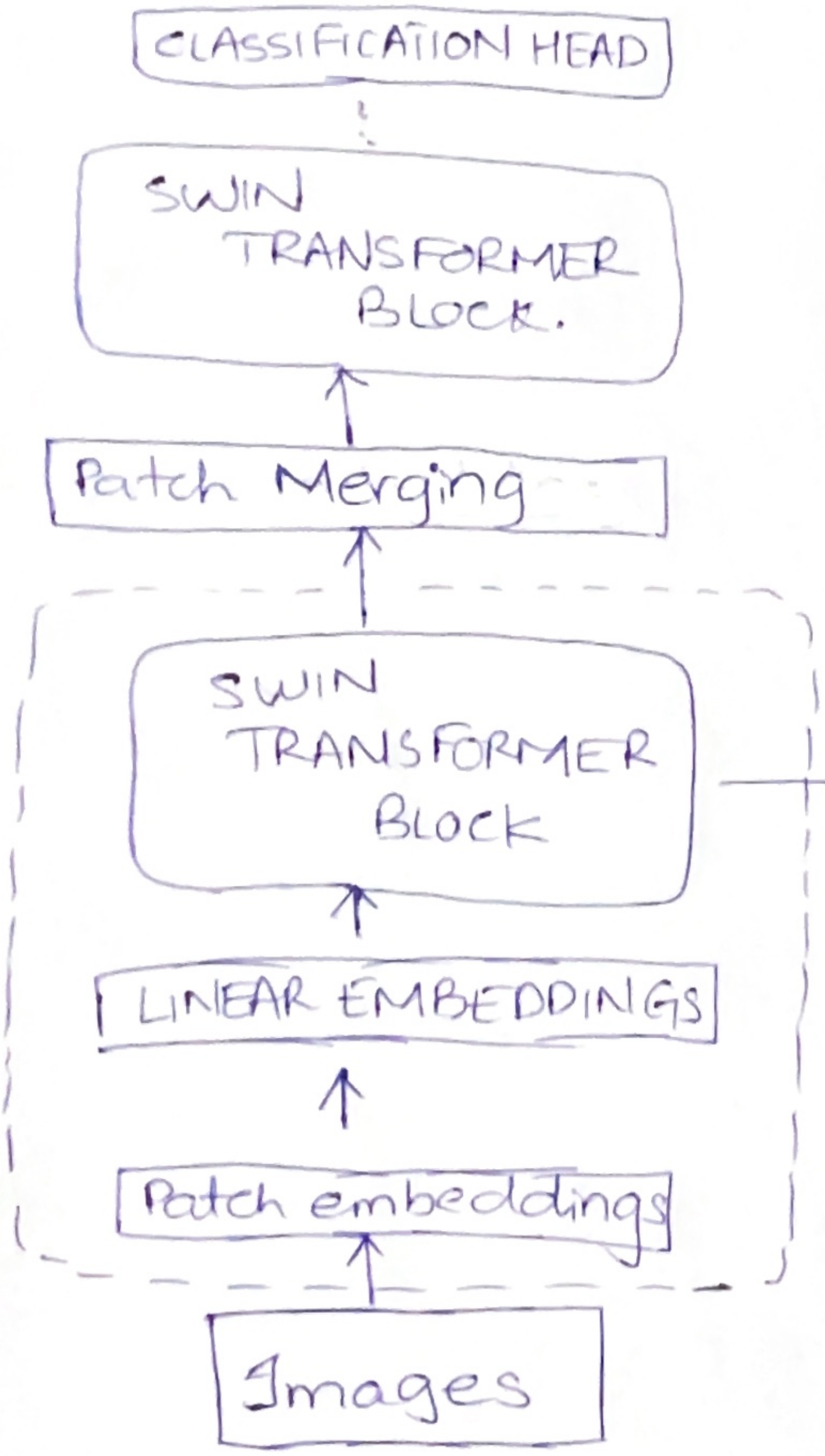
Windows will shift slightly e.g by (half the window size)

It means some patches from one window will now fall into a neighbour window.

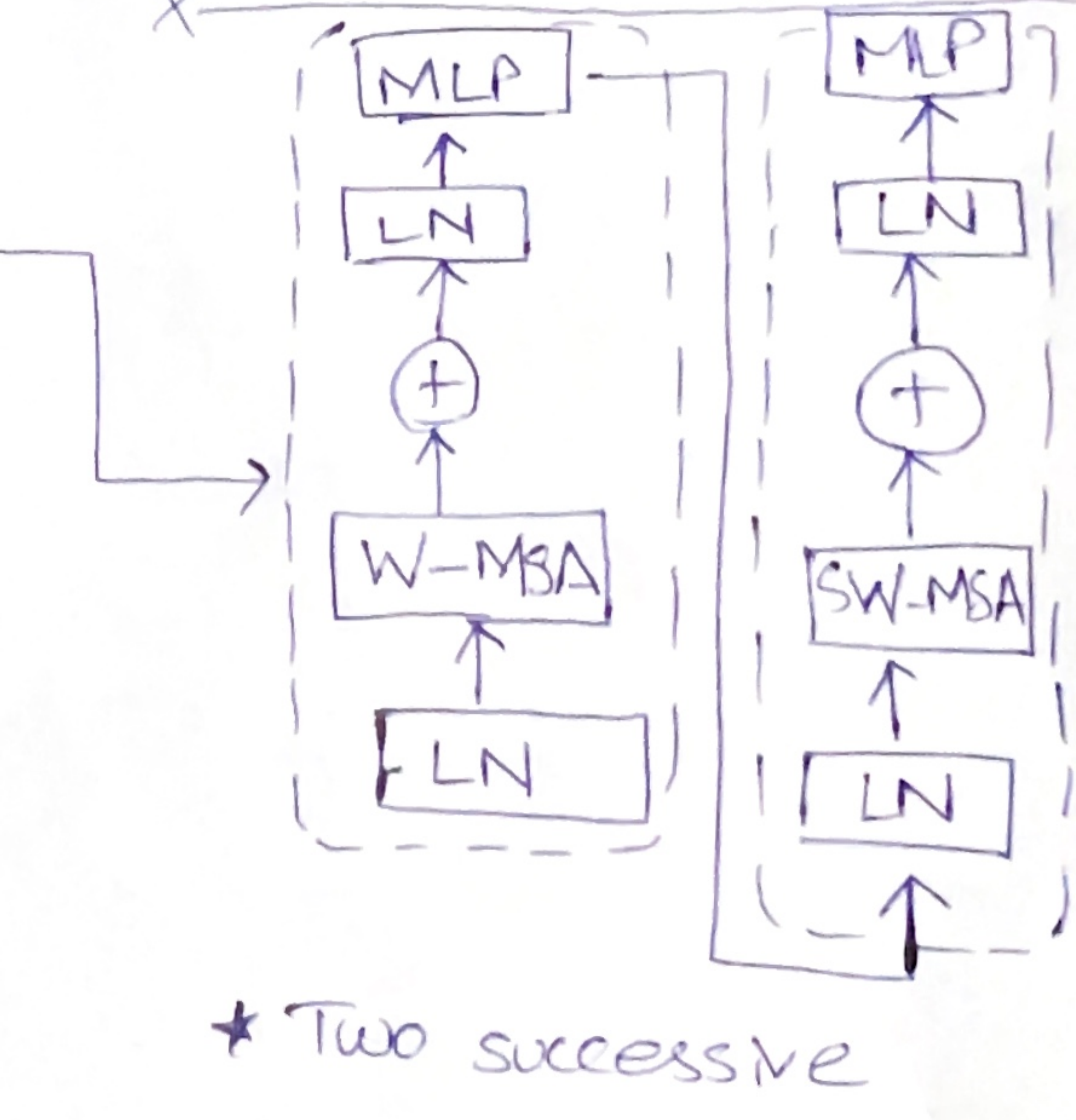
Now, patches from d/f windows can interact.

New windows are fermed & selfattention is applied again.

Transformer Archietecture.



* Basic SWINT Archietecture from original paper.



* Two successive SWIN Transfarmer blocks necessary.