**Transformers Architectures**

The transformer model is a machine learning model that has its origins in natural language processing (NLP) and is used to handle sequential input data. This data is typically a sequence of word embeddings. Transformers use a mechanism called attention to relate the sequential input to each other. As it comes from the NLP subfield of machine learning, it was first introduced as a way to relate words in translations from different words to one another. Thus, capturing what data points in the input had the largest effect on the output, i.e., what to pay attention to in a sense. When the concept is applied to machine vision the implementation is changed slightly, but the intuition behind it remains useful for understanding it. In machine vision applications it captures the relationship between pixels that are part of the same object. Specifically, transformers use multi-headed attention which applies the attention mechanism several times in parallel capturing different facets of the information and how different patches relate to one another.

**Vision Transformer (ViT)**

One of the first uses of the attention mechanism within the machine vision field was the Vision Transformer (ViT). It splits up an image into patches of pixels, grouped together in squares. Originally the patches were 16 by 16 pixels in size. These patches are embedded into a learned latent dimension and in turn, used analogously as word embeddings in the transformer models from the NLP field.

**Swin Transformer**

Shifted windows (Swin) Transformer is a transformer architecture developed by Microsoft and used for vision tasks. It was first proposed in March 2021 and established a new state of the art for several vision-related tasks, among them semantic segmentation where it achieved the highest metric on the ADE20K dataset. The Swin Transformer is a hierarchical transformer architecture which uses image patches of varying size in hierarchical layers where the goal is to capture both global and local features.

The model takes in an image which is split into several windows and processed by the first layer. The results of this is sent to the next layer which splits it up further. Since the window size between the layers vary while the number of patches in each window is constant, the patches get merged to fit inside the windows. In this architecture the layers with smaller window sizes are used to capture smaller features while the layers with larger window sizes are used to capture more global features. In the Swin Transformer architecture, attention is a key component and is computed inside each window. To propagate information between windows they are shifted so that they overlap. This is so that pixels that are part of the same object, but not inside the same window share information.

The use of windows in the image representation is the biggest difference between the Swin Transformer architecture and the ViT architecture is only split up into patches and not into windows of varying size.