Transformer-based Architectures for Image Features

Vision Transformer (ViT):

- ViT applies Transformer blocks to image patches instead of word tokens.
- Each block includes multi-head self-attention and a feed-forward network.
- ViT has shown competitive or superior performance compared to CNNs on many visual tasks.

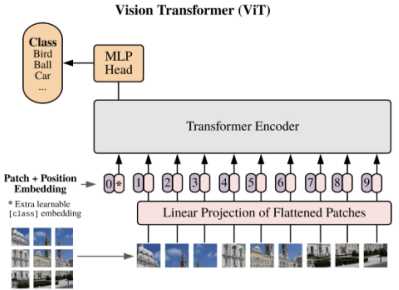

Figure 10: Vision transformer architecture