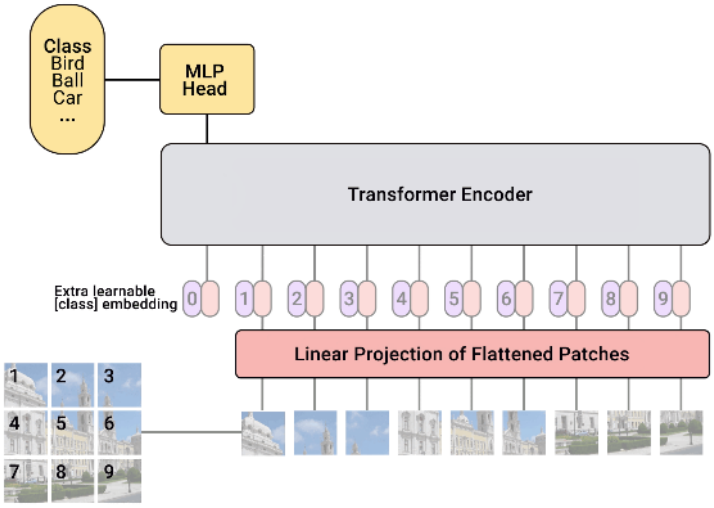
Visual Transformers (ViT)

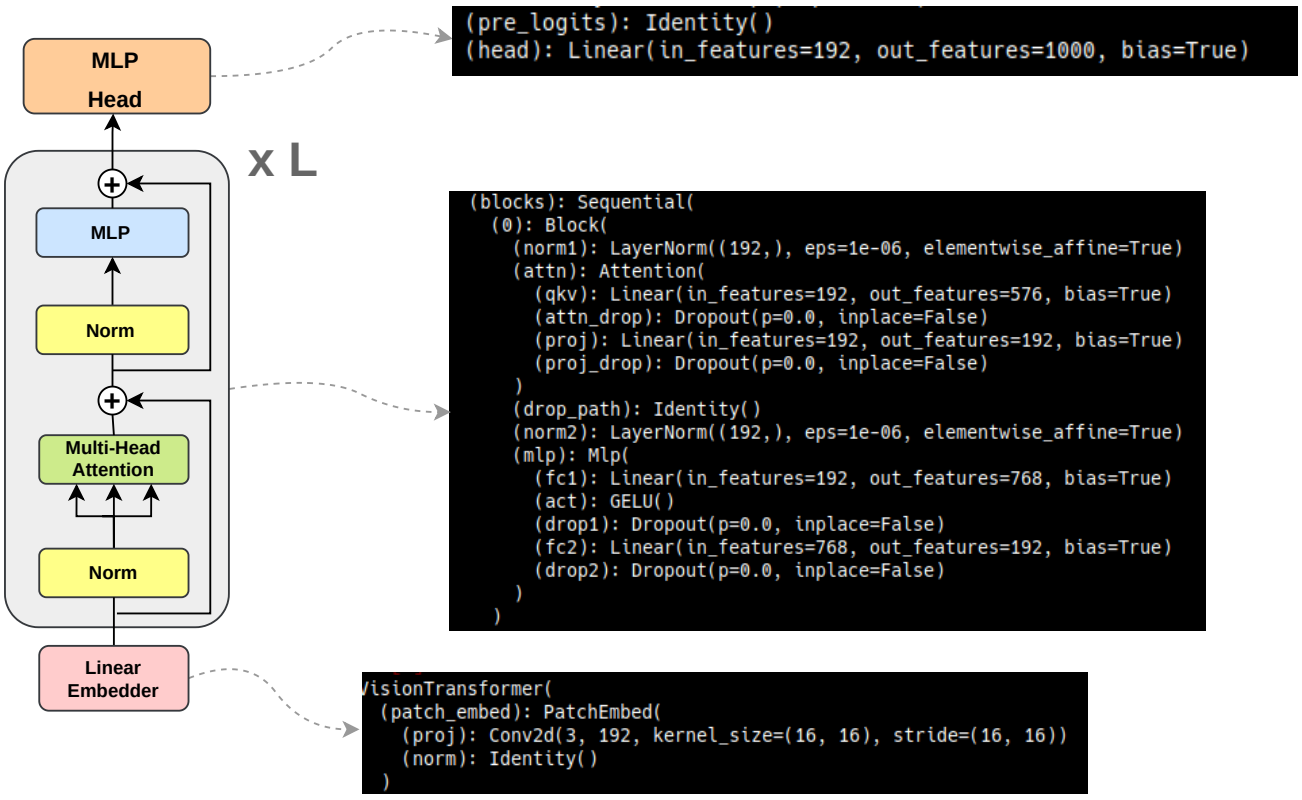
by Francesco Pelosin

We dissect, visually, how a Visual Transformer works. We will consider the ViT Tiny architecture, that is, a model composed by 12 layers each of them with 3 heads with embedding size of 192.

The model has been introduced in: https://arxiv.org/pdf/2012.12877.pdf and constitutes the "smallest" ViT architecture available. We consider the input images to be 224x224 pixels, with 3 channels and patch size of 16x16.

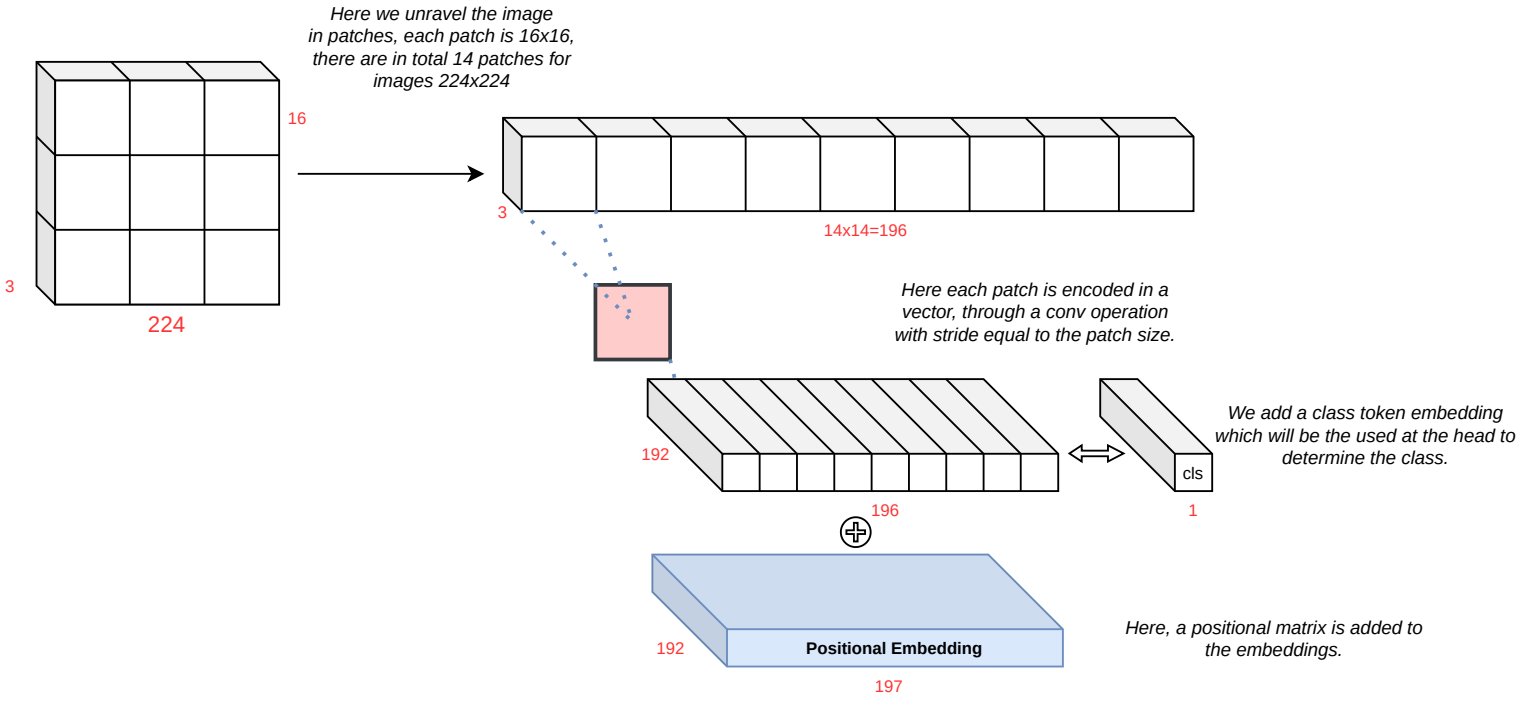
This is the general overview of a ViT. Each image is splitted in patches and then goes through a Linear Embedder. Here, we add a positional encoding and we append an extra [class] token which is the fundamental vector that is then used in the MLP Head to perform the classification.





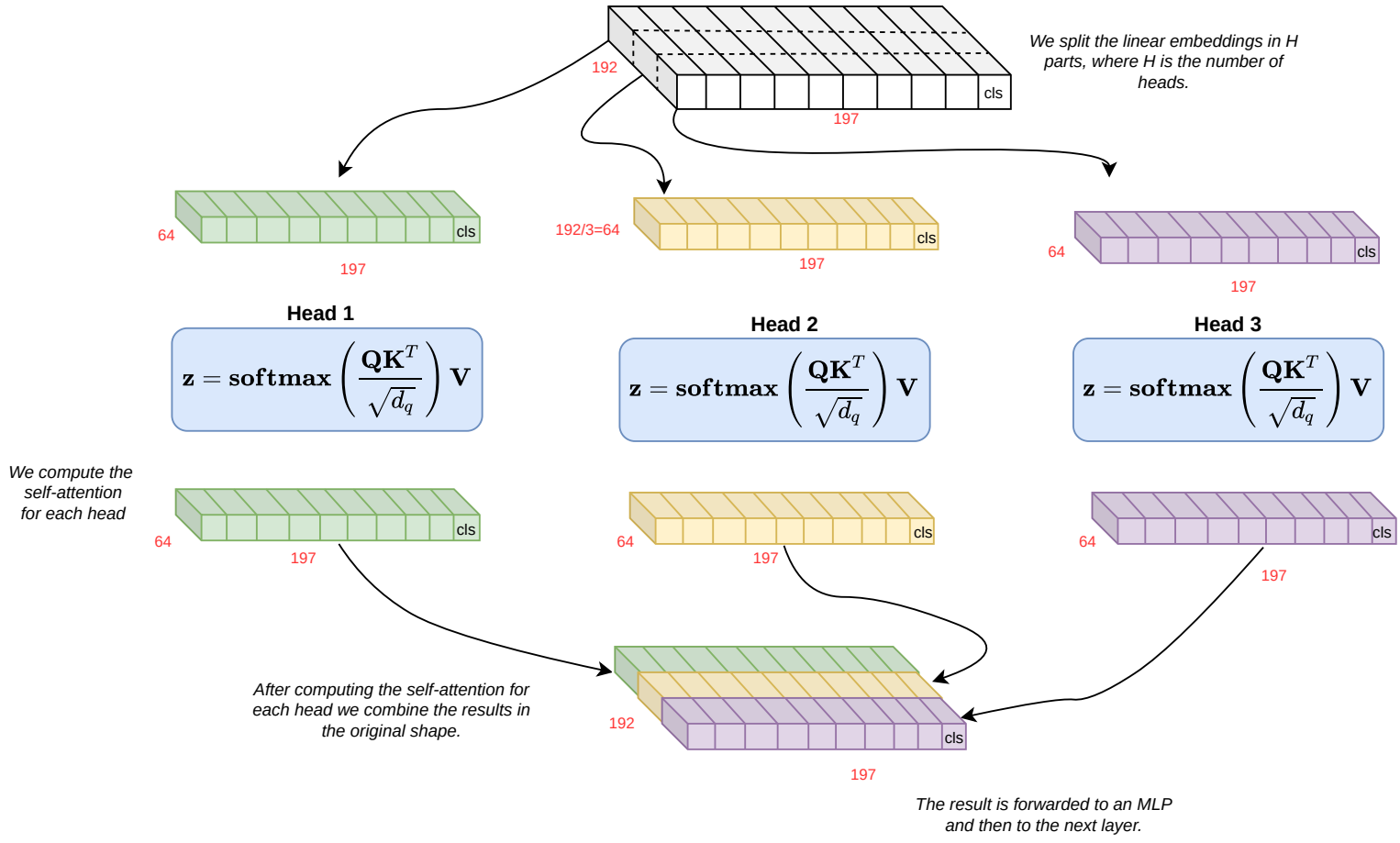
Linear Embedder

Linear Embedder



Multi-Head Attention

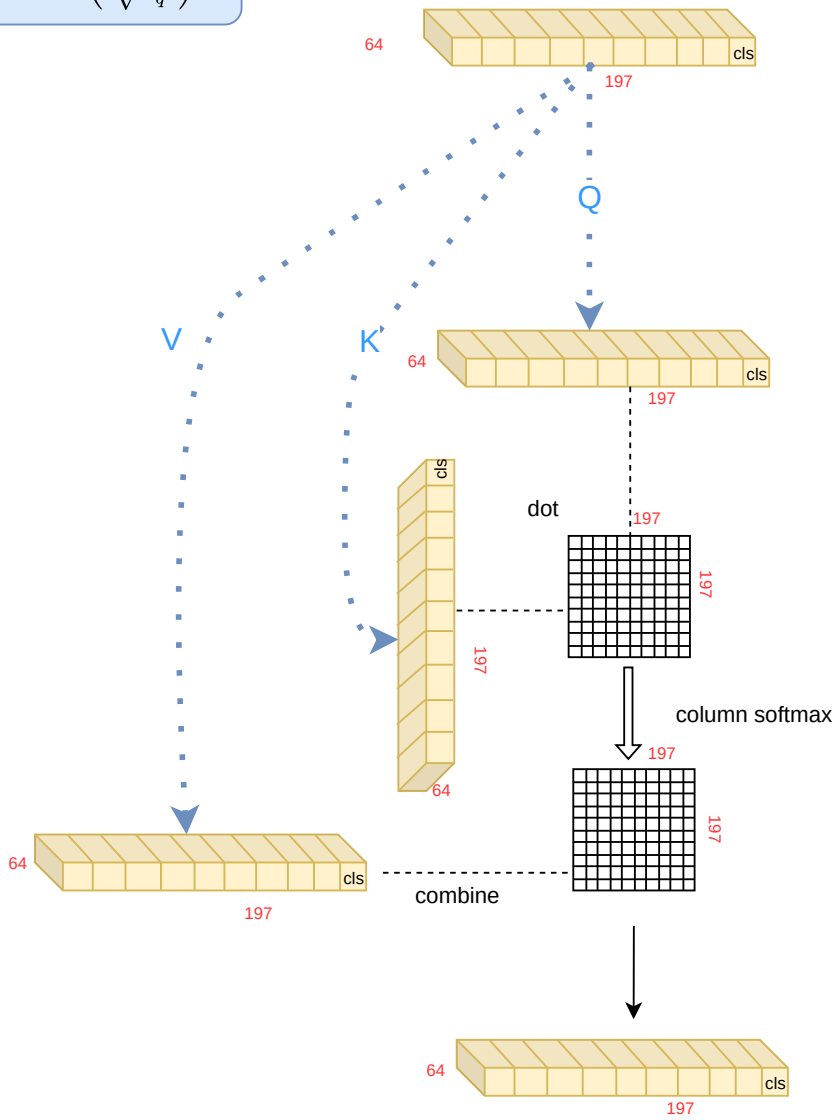
Multi-Head Attention



Attention Mechanism

$$\mathbf{z} = \mathbf{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_q}}\right)\mathbf{V}$$

Here we report the self attention mechanism that is performed for each head, in this case we have 3 heads. We report only Head 2.



MLP Head

MLP Head

At the last layer the MLP Head will only consider the cls token