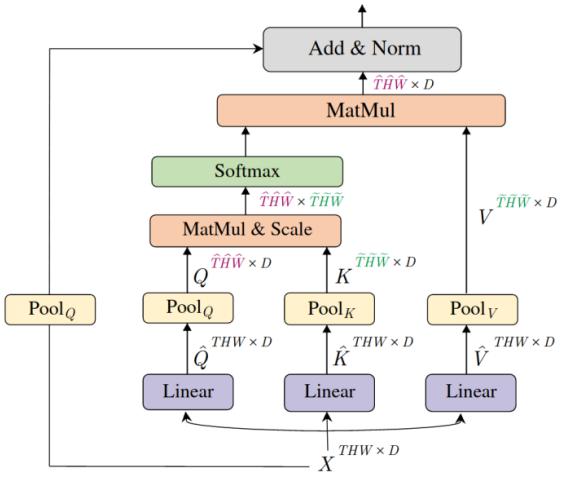
Multiscale Vision Transformers

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, Christoph Feichtenhofer

The main idea of MViT is to use a multi-scale approach across the spatiotemporal dimension (so the sequence length L = THW) and the channel dimension (D). Instead of having them fixed across the whole model, MViT starts with a large sequence length and a small channel dimension, then at the scaling stage, the keys, queries, and values are pooled across the sequence length to reduce the length of the output, while the channels dimension is slightly increased. The scaling is detailed in the table below.



stages	operators	output sizes
data layer	stride $\tau \times 1 \times 1$	$D \times T \times H \times W$
cube_1	$c_T \times c_H \times c_W$, D stride $s_T \times 4 \times 4$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale_2	$\begin{bmatrix} \text{MHPA}(D) \\ \text{MLP}(4D) \end{bmatrix} \times N_2$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale_3	$\begin{bmatrix} \text{MHPA}(2D) \\ \text{MLP}(8D) \end{bmatrix} \times N_3$	$2D \times \frac{T}{s_T} \times \frac{H}{8} \times \frac{W}{8}$
scale_4	$\begin{bmatrix} \text{MHPA}(4D) \\ \text{MLP}(16D) \end{bmatrix} \times N_4$	$4D \times \frac{T}{s_T} \times \frac{H}{16} \times \frac{W}{16}$
scale_5	$\begin{bmatrix} \text{MHPA}(8D) \\ \text{MLP}(32D) \end{bmatrix} \times N_5$	$8D \times \frac{T}{s_T} \times \frac{H}{32} \times \frac{W}{32}$

The scaling is done in stages, where the spatial dimensions are halved and the channel dimension is doubled at each scale, with a total of 4 scaling stages. Each stage is composed of a given number of transformer layers and the pooling is done only on the first layer of each stage. In this first layer, the scaling is done in both the main branch, but also the skip connection branch by adding the same pooling operation to the residual branch. Since the attention operation consists of (QK)V, the sequence length of the output depends only on the pooling done on the queries, so the paper proposes to have different pooling rates, one for the queries that control the overall sequence length, and one for (values, keys) that controls the computational cost of each attention operation, but both still vary comparably. While the scaling of the sequence length is done at the attention layer, the scaling of the channel dimension is done on the final MLP layer of the transformer layer. Another important aspect of MViT is the usage of a separate positional embedding in space & time instead of only a space embedding. So each patch of a given frame is encoded based on its position in the image and its position in the clip, where all patches of the same image will share the same temporal embedding, and all of the patches of the same spatial location will share the same spatial embedding.