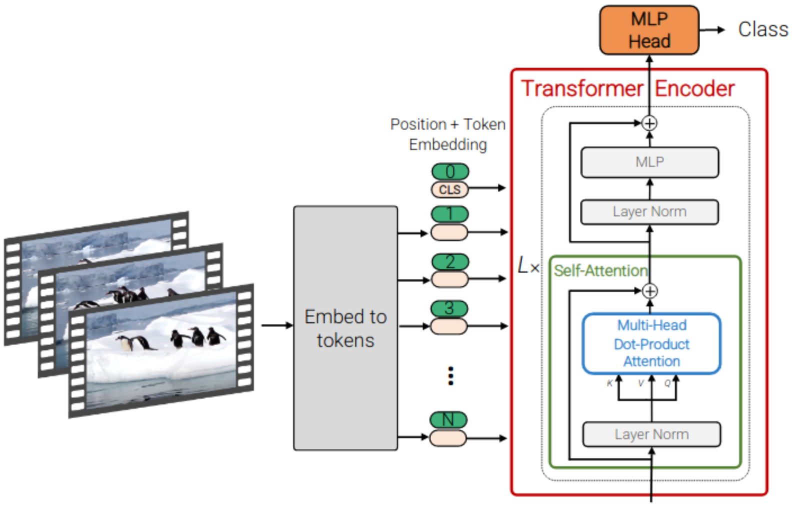
ViViT: A Video Vision Transformer

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid



Similar to TimeFormer, ViViT tests many different variants of the attention layers in order to consider both the spatial and temporal contents of the inputs. Additionally, they also propose a new method of embedding the inputs, where the embedding is done both spatially and temporally instead of per-frame basis with some hand-designed regularization and initialization schemes to overcome the problems of data scarcity.

Embedding: the paper proposes two types of embedding,

- *2D Embedding*; which is the standard per frame basis where each image is converted into N patches, then aggregated across time with T frames, resulting in an input of N x T steps.
- *Tubelet embedding:* which consists of applying a 3D convolution over the T x H x W inputs with a stride equaling the size of the kernel to avoid having any overlap, where each input token corresponds to one application of the 3D convolutions across the input.

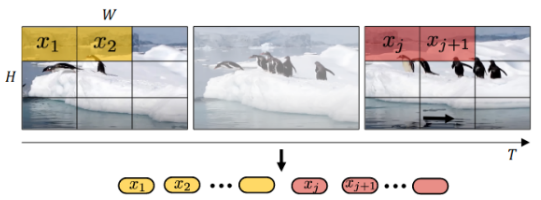


Figure 2: Uniform frame sampling: We simply sample n_t frames, and embed each 2D frame independently following ViT [18].

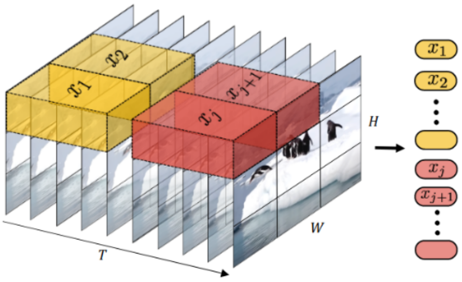
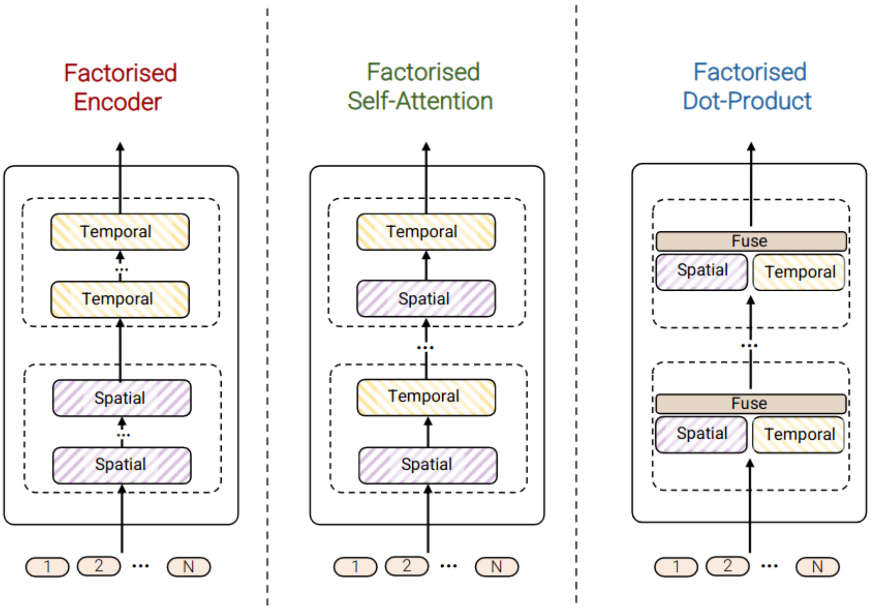


Figure 3: Tubelet embedding. We extract and linearly embed non-overlapping tubelets that span the spatio-temporal input volume.

Models: In addition to a baseline of spatiotemporal with one attention across all dimensions, temporal and spatial, the paper proposes three of models based on three types of spatiotemporal attention layers.

- *Factorised Encoder*: this is similar to VTN & STAM, where we start with spatial layers and then stack temporal layers on top.
- *Factorised Self-Attention*: this is similar to TimeFormer, where each attention layer has a spatial attention layer followed by a temporal one.
- *Factorised Dot-Product*: this consists of applying both the spatial and temporal attention in parallel, and the two outputs are then concatenated and followed by an MLP projection.



Initialization: one of the important aspects of ViViT is the initialization process. The initialization is based on an image-based pre-trained transformer (on ImageNet 21M for example), first, for the positional embeddings, in the 2D version, they simply copy and repeat the single embedding T times. While for tubelet embedding, the embedding layer consists of 3D convolutions instead of 2D, so they either inflated the 2D into a 3D or simply initialize the center kernel with the 2D and the rest as zero (the second one, called Central frame init, is better). As for the attention layers, the spatial attention is initialized from the image model while the temporal ones are initialized as zeros, so at the start, the model is just a spatial transformer with no temporal extraction.

The results show that Tubelet embedding with Central frame initialization and factorized encoder (first spatial encoder then temporal encoder, as in STAM & VTN) is better.