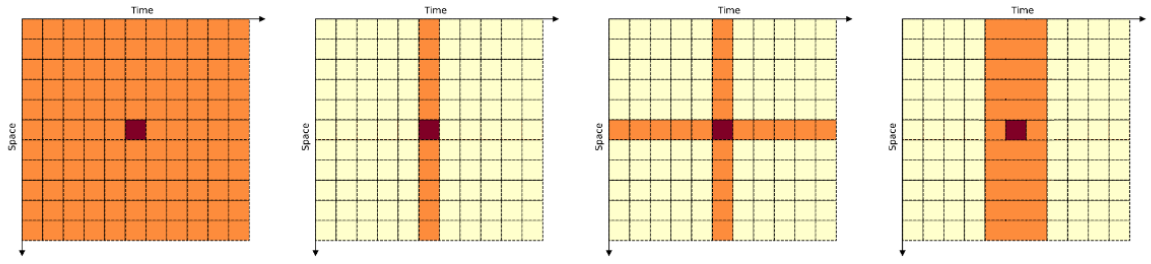
# Space-time Mixing Attention for Video Transformer

Adrian Bulat, Juan-Manuel Perez-Rua, Swathikiran Sudhakaran, Brais Martinez, Georgios Tzimiropoulos

X-ViT instead of factorizing the spatiotemporal attention into two disjoint ones, the paper proposes to define a better joint attention computation to reduce the quadratic computation (THW x THW) of the spatiotemporal attention operation. Another alternative used by previous models is spatial only (HW x HW), factorized (TxT + HW x HW), or pooled (THW/scale x THW/scale). X-ViT proposes to use a temporal window, where the patch of each image attends only to frames within the predefined temporal window. This way, the complexity of the attention mechanism depends quadratically only on the spatial scales while being linearly dependent on the temporal scale. Additionally, the temporal receptive field grows linearly with the number of layers resulting in a larger temporal context.




(a) Full space-time attention: $O(T^2S^2)$ (b) Spatial-only attention: $O(TS^2)$ (c) TimeSformer [3]: $O(T^2S + TS^2)$ (d) Ours: $O(TS^2)$

Figure 1: Different approaches to space-time self-attention for video recognition. In all cases, the key locations that the query vector, located at the center of the grid in red, attends are shown in orange. Unlike prior work, our key vector is constructed by mixing information from tokens located at the same spatial location within a local temporal window. Our method then performs self-attention with these tokens. Note that our mechanism allows for an efficient approximation of local space-time attention at no extra cost when compared to a spatial-only attention model.

For a window of size W, the standard implementation will give complexity of (W x T x HW x HW). However, it can be further reduced, where instead of computing the keys W times at each temporal location, such keys were already computed T times. So instead of recomputing them TW times, we can simply shift (W/2 to the right and left) and duplicate the original T keys resulting in TW keys with only T computations.

Since the model operates directly on all frames, we end up with T per frame CLS embeddings, and to aggregate them, the paper proposes final temporal attention over the T tokens (TA). Additionally, another addition the paper proposes is summary tokens, where before computing the keys and values, a 1-D vector summary of the same size the heads computed as the average pooling overall spatial locations of each one of the frames is appended to them before the attention computation.