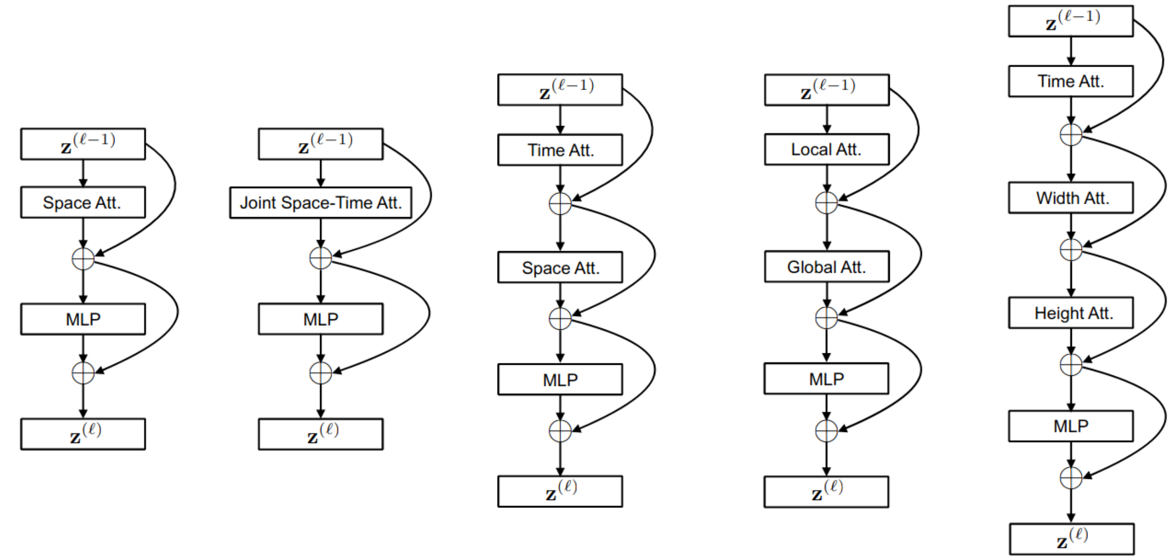
Is Space-Time Attention All You Need for Video Understanding?

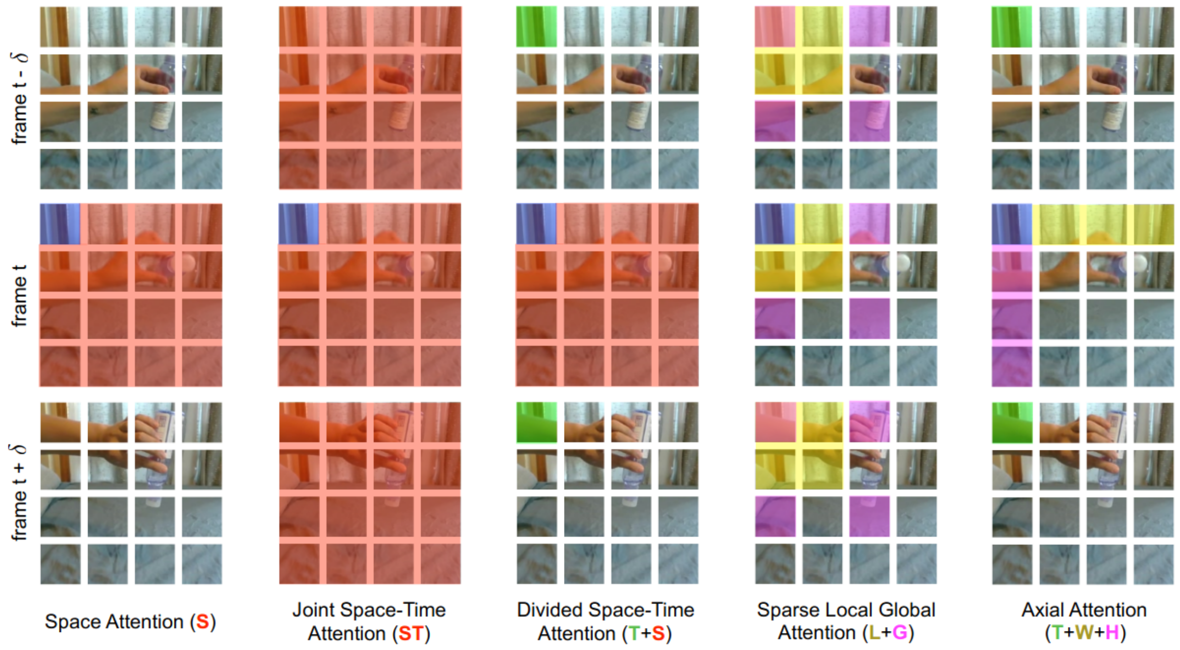
Gedas Bertasius, Heng Wang, Lorenzo Torresani



Space Attention (S) | Joint Space-Time Attention (ST) | Divided Space-Time Attention (T+S) | Sparse Local Global Attention (L+G) | Axial Attention (T+W+H)

Instead of having a separated spatial and temporal encoder, TimeFormer tries to have both a spatial and a temporal component at each transformer layer. For this, they propose different types of attention:

- Space only: each frame only attends to itself.
- Joint space-time: each frame attends to all of the rest.
- Divided space-time: first attend spatially, then temporally, at each transformer layer.
- Sparse: attend both spatially and temporally to a subset of all the patches of all frames.
- Axial: first attend to patches of the same row and column of the same frame, then temporally to all of the patches of the same location across time.


frame t - δ
frame t
frame t + δ
Space Attention (S)
Joint Space-Time Attention (ST)
Divided Space-Time Attention (T+S)
Sparse Local Global Attention (L+G)
Axial Attention (T+W+H)

By testing these types of attention, the authors find that divided space-time performs best and on par with joint attention, while being more memory efficient, and they call the resulting model TimeFormer. Additionally, they show that TimeFormer can be quite robust to the choice of both the number of input frames and the number of input clips used during testing since it is able to span the whole video with a single input clip with a large sampling stride.