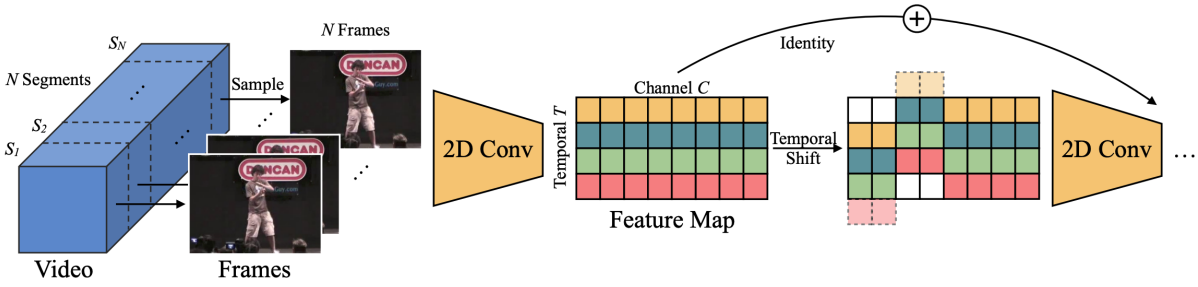
TSM: Temporal Shift Module for Efficient Video Understanding

Ji Lin, Chuang Gan, Song Han

The widely used and standard 2D CNNs are quite efficient and can yield good results on video understanding tasks, however, they lack the temporal information by only operating on individual frames, which decreases their performances. On the other hand, 3D CNNs are capable of jointly learning both spatial and temporal features, making them better at the task of action recognition, but they remain computationally expensive.



TSM module tries to solve this by shifting some of the channels of the input features in the temporal dimension while still maintaining the usage 2D CNNs, this way, the new activation computed over each temporal dimension separately will contain information from different frames, depending on the size of the shift. However, the shift operation, while having zero computational flop, is quite memory heavy since the activations need to be moved from one location to the other. Additionally, shifting too many channels can hurt the performances since the spatial modeling will be incoherent. To solve these issues, TSM shits only small subsets of the channels across the temporal dimension (bidirectionally for offline videos and unidirectionally for online video), in addition to adding the TSM module inside a residual branch (the output is the original input plus the shifted and convolved input) to maintain the activations of the current frame while inserting some temporal information.

Network: Given a video, it is divided into segments and then a number of frames are sampled equally from each segment, and each frame is processed with a 2D CNN, but with TSM modules at each residual block of the network. This way, at each block, the model will have access to 3 frames or two additional ones since there is one shit inward in time and one backward in time.