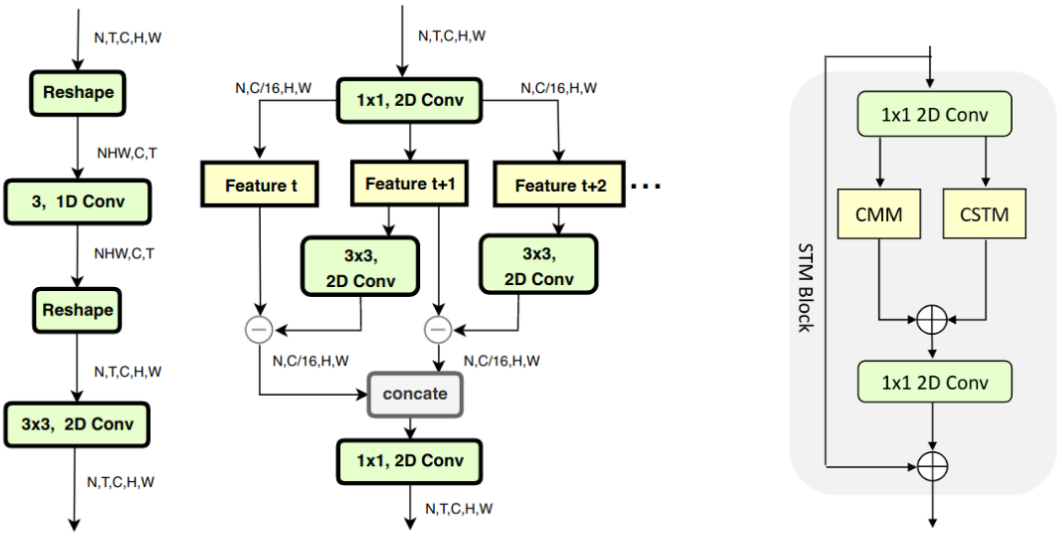
STM: SpatioTemporal and Motion Encoding for Action Recognition

Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, Junjie Yan

The two dominant approaches for video recognition is 2D CNNs with two-stream, an appearance stream operating over individual RGB frame, modeling only spatial information, and a motion stream modeling temporal information by taking as input the optical flows. And the second method is 3D CNNs that model both using 3D convolutions. However, with the first method the optical flows only model recursive motion and lacks long-term temporal and spatial interaction, while 3D CNNs still require a second stream with explicit motion information. To solve this, the paper proposes STM Networks that integrate both spatiotemporal features with motion information.

The STM network consists of adding two modules to the 2D CNNs, a Channel-wise SpatioTemporal Module (CSTM) module to inject temporal information into the spatial features, and Channel-wise Motion Module (CMM) to model the motion explicitly without the need for optical flows as an additional input.



(a) CSTM

(b) CMM

The CSTM module consists of reshaping the input tenors from N x T x C x H x W into NHW x C x T, and then applying a size 3 1D convolution over the temporal dimension. Since each channel has different spatial features, each channel has a specific filter, and this can be implemented by a group-wise 1D conv. The final CSTM module then consists of a reshape, a channel-wise temporal conv, a reshape into the original format, and a 3x3 2D convolution. The CMM module of a first 1x1 conv to reduce the channel dimension, then for each frame at time t, the two neighboring frames at t-1 and t+1 are extracted, and an element-wise subtraction between each two (t-1;t and t;t+1) is computed and then results are concatenated to model motion between the frames. Finally, the two modules are applied in parallel in between the 1x1 convs as in the standard resnet block; resulting in an STM block used to design the model.

TEA (Temporal Excitation and Aggregation)

A very similar model to CSN is TEA, which also proposes two modules, a motion excitation (ME) module, and a multiple temporal aggregation (MTA) module, specifically designed to capture both short- and long-range temporal evolution. The ME module calculates the feature-level temporal differences from spatiotemporal features. It then utilizes the differences to excite the motion-sensitive channels of the features. The MTA module proposes to deform the local convolution to a group of sub- convolutions, forming a hierarchical residual architecture. The TEA block uses these two modules in series in between two 1x1 2D convs.

MotionSqueeze (MotionSqueeze: Neural Motion Feature Learning for Video Understanding)

Another work that tries to model the motion by defining a new module that operated over adjacent temporal features is MotionSqueeze Net. The introduced module first computes the correlation between adjacent frames of the features of two given pixels with some displacement between them, where for each pixel in frame t we consider all possible displacement in a window of size K x K in frame t+1. Then one displacement is chosen from the K x K possible ones, but to make it differentiable, they use a weighted average over all possible K x K values; this is followed by a Gaussian smoothing to remove outliers.