Temporal Pyramid Network for Action Recognition

Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, Bolei Zhou

Instead of constructing multiple level of features with different temporal scales using multi-branch networks (like slowfast), TPN tries to reduce the computational overhead and reuse the produced features of the backbone at different blocks with a single and temporally fixed input size.

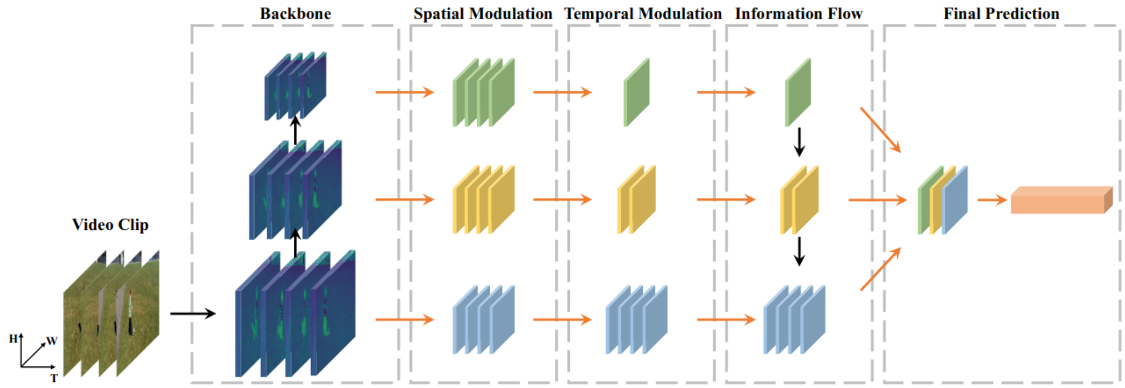

Figure 2. **Framework of TPN:** *Backbone Network* to extract multiple level features. *Spatial Semantic Modulation* spatially downsamples features to align semantics. *Temporal Rate Modulation* temporally downsamples features to adjust relative tempo among levels. *Information Flow* aggregates features in various directions to enhance and enrich level-wise representations. *Final Prediction* rescales and concatenates all levels of pyramid along channel dimension. Note that the channel dimensions in *Final Prediction* and corresponding operations are omitted for brevity.

TPN first select a set of blocks to extract the features from, such blocks comes from different levels of the network with different temporal receptive fields making them suitable from modeling the input at different temporal scales. First each level is spatially processed to align spatial semantics of the features using a stack of convolutions. Then a temporal modulation to control the temporal flexibility, with a temporal convolution and a temporal pooling. The two are then fused, either in both direction or from spatial to temporal. For Up to bottom, each time a spatiotemporal downsampling is applied to adjust the T x H x W scales to the lower volume, as for bottom to top, an nearest upsampling is performed at each step.