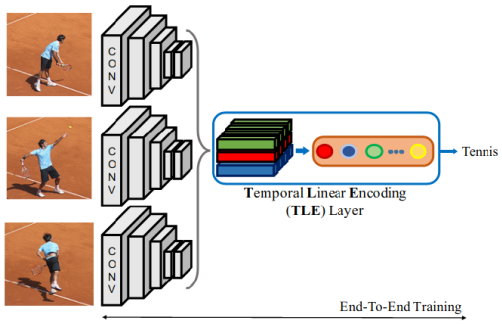
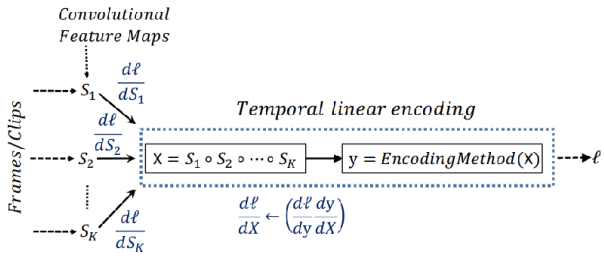
Deep Temporal Linear Encoding Networks

Ali Diba, Vivek Sharma, Luc Van Gool

The main component of the proposed models is the TLE module, a new spatio-temporal feature encoding module that aggregates multiple video segments over long time ranges. Inspired by the success of IDTs, which showed that tracking densely sampled point over the frames with optical flow can give good video representations. TLE consists of efficiently aggregating the frames together on a single feature space to encode all of the clips or videos, instead of using late fusion where each frame or clip is scored separately and the scores are then aggregated.





Given K features maps generated by a CNN over K (here K = 3) input segments, where each feature map is of size H x W x C. The paper investigates different linear aggregation methods to encoder all of the K segments into a compact and robust representation:

- Element-wise average.
- Element-wise maximum.
- Element-wise multiplication. The best is element-wise multiplication which was therefore selected.

After the aggregation, the resulting features are then encoded, producing a linearly encoded feature vector where every channel of the aggregated features interact with the rest of the channels, resulting in better representations. For such an encoding, the paper investigates the following:

- Bilinear Models: compute the outer product between the aggregated features and themselves, resulting in features of size C^2, then followed by a projection into a lower dimensional space using a linear layer.
- Fully-connected pooling: The aggregated features are directly fed into a linear layer followed by a classification layer. The best is bilinear aggregation method.

The final network: a two-stream network with spatial and temporal networks. The spatial net operates on RGB frames, and the temporal net operates on a stack of 10 optical flow frames. The model can also be 3D based, this time the input are 3 clips for video recognition, instead of 3 frame for clip recognition. The TLE module is added before the classification layer to aggregate the K feature maps followed by the classification layer. For a 2 stream network, TLE is added on both nets followed by a score averaging.