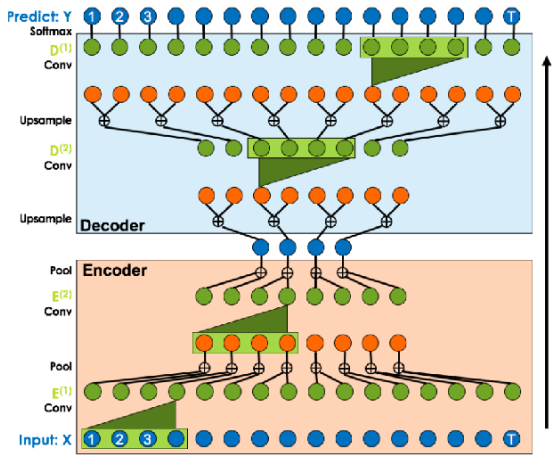
Temporal Convolutional Networks for Action Segmentation and Detection

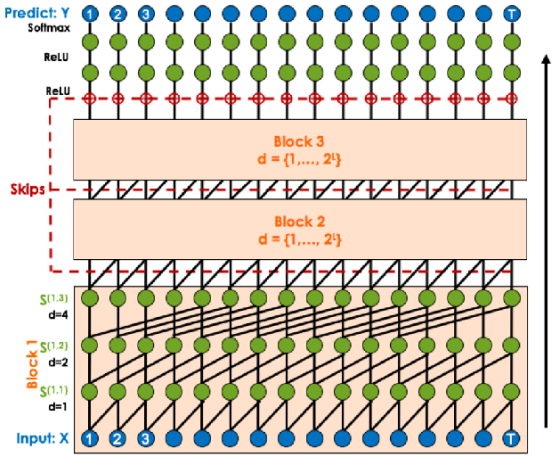
Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, Gregory D. Hager

Temporal Convolutional Networks (TCNs) use a hierarchy of temporal convolutions to perform fine-grained action segmentation or detection. Instead of extracting the spatio-temporal features from video frames and then feeding them to a temporal classifier. With TCNs, we have the following: 1) computations are performed layer-wise, meaning every time-step is updated simultaneously instead of updating sequentially per-frame (2) convolutions are computed across time, and (3) predictions at each frame are a function of a fixed-length period of time (temporal receptive field). Based on such properties, the paper proposes two architectures:

- Encoder-Decoder TCN: consisting of L encoding layers, where each layer contains a temporal convolution, a non-linear activation and max pooling across time that halves the number of temporal steps. The decoder is similar to the encoder except that the max pooling layers are replace with up-sampling layers that simply duplicate the entries temporally.
- Dilated TCN: follows a similar structure to that of WaveNet, but adapted for action segmentation. It consists of a series of convolutional layer with the same number of filter and with dilated convolutions with increasing rate with higher layers. Each layer consists of dilated convolution, a non-linear activation and a residual connection that combines the layer's input and its convolved version. The convolution consists of W1 x F(t-s) + W2 x F(t). With s as the dilation rate and F as the input features, followed by skip connection.



Encoder-decoder TCN



Dilated TCN

Note that for both models, there is a causal version where the prediction at each frame only depend on the previous ones, here convolution of the encoder-decoder network is restricted to only the *d* previous frames. Or acausal where we have access to both the forward and backward directions. Here, in the

dilated version, the input to each convolution takes the center, left with a given dilation step, and right with a given dilation step.

Other variations:

- MS-TCN or Multi-Stage TCN: Stacks multiple Dilated TCNs on top of each other for better refinement. In addition to a smoothing loss to reduce the over-segmentation found in some video predictions, which is simply an MSE loss between two class probabilities at time steps t and t-1. The final loss is then the sum of the losses at each stage.
- MS-TCN++: Here, the architecture is the same as MS-TCN but where the first few stages have a modified version of the dilated convolutions. The new version is called dual dilated convolutions where the output is the concatenation of the output of two dilated convolution with different dilation rates. The dilation rate of first one decreases with higher layers while the second increases.

Important Note: All of these models take as input 1D vector which are I3D features for each video frame, so they don't take as input the raw frames.