Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

Joao Carreira, Andrew Zisserman

Given how small the previously dominant datasets (UCF-101 and HMDB-51), it can be quite difficult to evaluate the proposed architectures and find the best design choices. To solve this, the paper investigates the dominant design choices using Kinetics dataset, which has two orders of magnitude more data, with 400 human action classes and over 400 clips per class. The comparison will also be based on a new version of 3D convolutions called inflated 3D convolutions, where 2D ImagetNet pertained convolutions are inflated over the time dimension, transforming them into 3D convolution adapted for video understanding tasks.

Inflating 2D convs: each 2D filter can be converted into a 3D one by simply adding the temporal dimension, going from N x N to N x N x N. Where the N x N original spatial weights are first copied N times, and then divided by N to ensure that the new filter gives the same response as the original one.

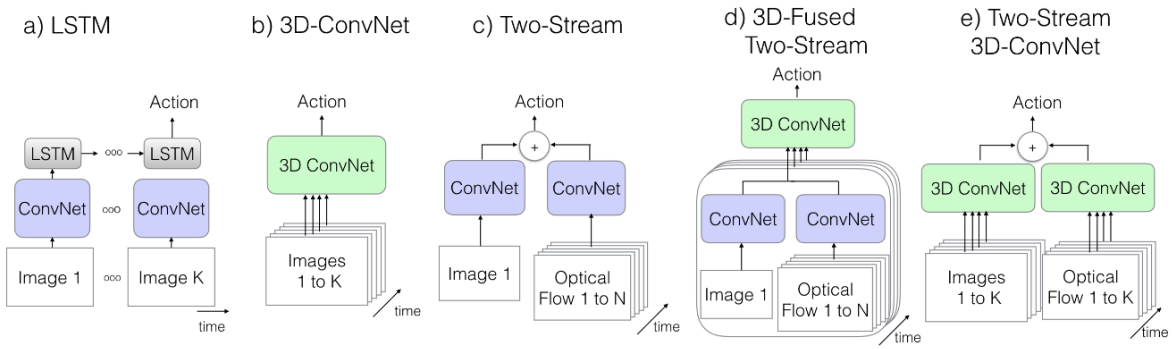

Figure 2. Video architectures considered in this paper. K stands for the total number of frames in a video, whereas N stands for a subset of neighboring frames of the video.

The authors compare old methods, mainly:

- ConvNet+LSTM where 2D convolutions are applied over K frames then processed temporally with an LSTM.
- 3D ConvNets process the input frames both spatially and temporally using 3D convolutions.
- Two-Stream Networks: that uses optical flows in addition to RGB frames in a two-stream design for an enhanced motion and a recursive motion understanding, where both inputs are processed using 2D convolutions.
- Two-Stream Inflated 3D ConvNets: where both streams are processed using 3D convolution. In all of the above, the 3D convolutions can be replaced with the new 3D inflated convolutions for better

results.

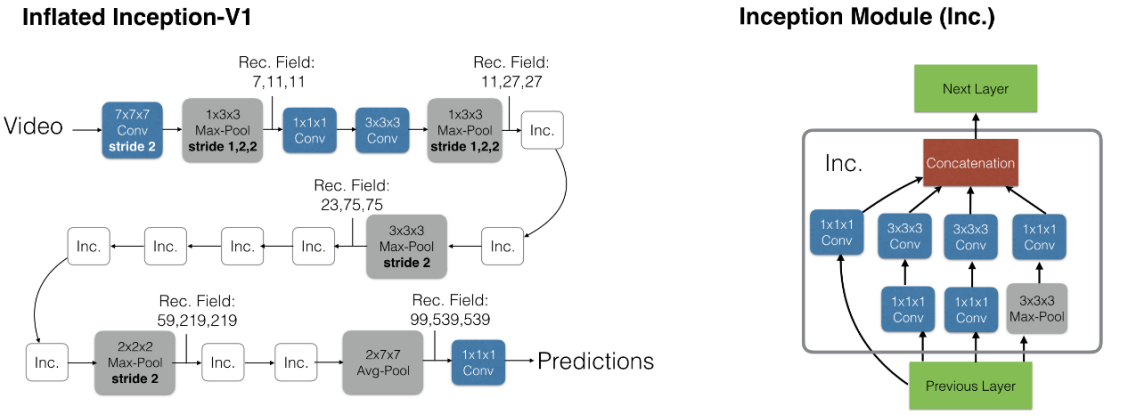

Figure 3. The Inflated Inception-V1 architecture (left) and its detailed inception submodule (right). The strides of convolution and pooling operators are 1 where not specified, and batch normalization layers, ReLu's and the softmax at the end are not shown. The theoretical sizes of receptive field sizes for a few layers in the network are provided in the format "time,x,y" – the units are frames and pixels. The predictions are obtained convolutionally in time and averaged.

Network: Given the new inflated 3D, any pre-existing architecture can be converted into a 3D in a straightforward manner. However, in 2D networks, when doing a pooling operation, the pooling kernels and strides are the same in both spatial directions. However, this is not the case with the time domain, where such parameters need to depend on the frame rate, and if the temporal receptive field grows too quickly and in tandem with the spatial receptive field, we might conflate edges from different objects if it grows too fast, and if it grows too slowly, it may not capture scene dynamics well. To strike a good balance, the proposed I3D, a 3D version of inception, does not have any temporal pooling in the first two max-pooling layers, while having symmetric temporal and spatial parameters for the rest, except the last one where the average pooling is applied over 2 frames.