Spatiotemporal Residual Networks for Video Action Recognition

Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes

ST-ResNet (Spatiotemporal Residual Nets): The model builds on the two steam approach (temporal with optical flows as inputs, and spatial with RGB images as inputs) with a final score fusion. However, the ST-RestNet model is initialized from ResNet trained on ImageNet, where the 2D convs are transformed from 2D to 3D by inflating them as in I3D, . The model then becomes a 3D convolutional 2-stream net. Additionally, instead of using a late fusion of the two-stream, ST-RestNet adds a residual connection between the two streams and jointly fine-tuning them.

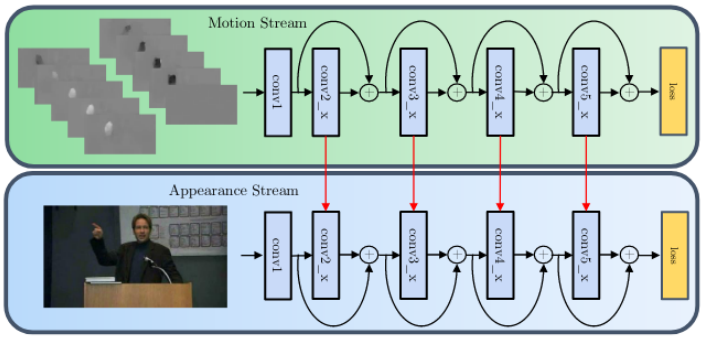

Figure 1: Our method introduces residual connections in a two-stream ConvNet model [20]. The two networks separately capture spatial (appearance) and temporal (motion) information to recognize the input sequences. We do not use residuals from the spatial into the temporal stream as this would bias both losses towards appearance information.

STM-ResNet (Spatiotemporal Multiplier Nets): Instead of a simple sum-based residual fusion as in ST-ResNet, this paper tries to investigate different fusion methods in a more systematic way, in order to find the best fusion method with such residual connections. The residuals can be additive or multiplicative, and one way (motion to appearance or vice-versa) or two-way interaction. The authors then found that the best ones are motion-to-appearance one-way connections, where the multiplicative connection performs a bit better. The final model then consists of 4 residual blocks, each one with three convolutions, where the residual connections are added after the first convs for each one of the 4 residual blocks.

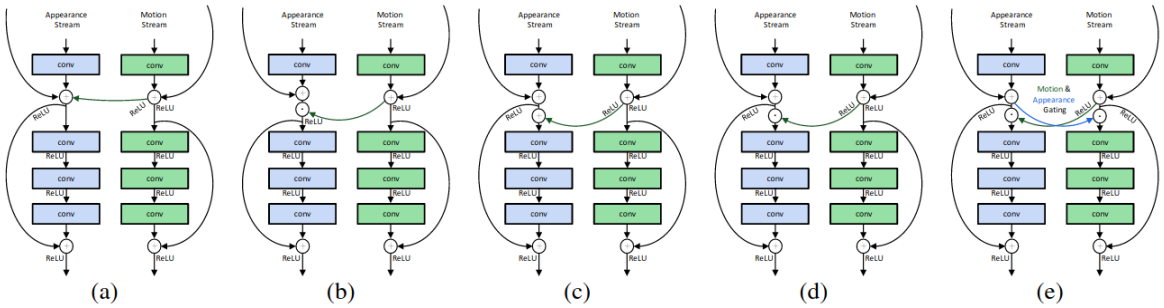

Figure 2. Different types of motion interactions between the two streams enables the learning of local spatiotemporal features. (a)-(d) show unidirectional connections from the motion into the appearance path and (e) illustrates bidirectional gating across streams.