**TIMECEPTION**

FOR COMPLEX ACTION RECOGNITION

NOURELDIEN HUSSEIN, EFSTRATIOS GAVVES, ARNOLD SMEULDERS

**PROBLEM**

Complex actions of Charades are 3D scenes, compared to 5 sec of Kinetics.

**EXTENT**

One-action, comprising complex action, vary in their temporal extents.

**DEPENDENCY**

Temporal dependency, albeit weak, between the one-actions.

**METHOD**

**TEMPORAL-ONLY CONV**

Depthwise separable 1D conv to reduce complexity of 3D conv from $O(t \cdot c^3)$ to $O(t \cdot c)$.

**MULTI-SCALE KERNELS**

Different kernel sizes ($k$) or dilation rates ($d$) to account for varieties in temporal extents of one-actions.

**EFFICIENT MODULAR LAYER**

Grouped conv and concat+shuffle to reduce the computational cost of typical 3D conv.

**RESULTS**

**DATASET: CHARADES**

Improve over I3D, R3D, NL and GCN with much less parameters. Temporal footprint is 10-fold longer than our non-local. Computational cost is much less than related works.

**LAYER EFFECTIVENESS**

Timeception monotonically improves as the network goes deeper. The same result is confirmed when using different backbones, as ResNet and I3D.

**MULTI-SCALE KERNELS**

Convolutions with multi-scale kernels outperform their fixed-sized counterparts. Performance of different dilation rates ($d$) is comparable with that of different kernel sizes ($k$).

**LONG-RANGE DEPENDENCY**

For complex actions, Timeception does better than related methods in modeling the long-range temporal dependencies. But for some short-range, simple actions, it is outperformed.