UNIVERSITY OF AMSTERDAM

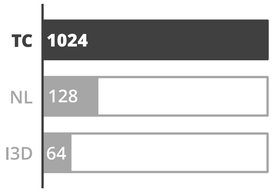
ORAL PRESENTATION

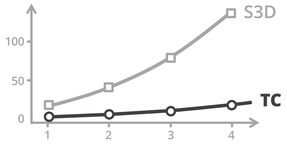
TIMECEPTION

FOR COMPLEX ACTION RECOGNITION

NOURELDIEN HUSSEIN, EFSTRATIOS GAVVES, ARNOLD SMEULDERS

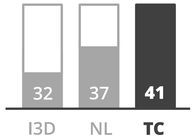
10x frames





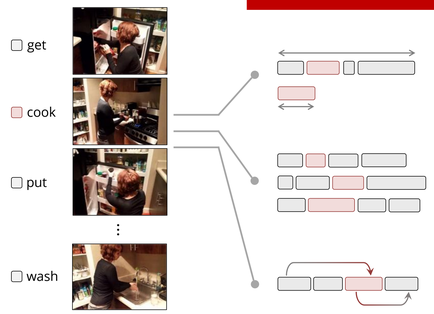
2.8m params

3.6% improve



CVPR 2019
JUNE 16-20, LONG BEACH, CA

PROBLEM



Complex Action: Cooking a Meal

LONG-RANGE

Complex actions of Charades are 30 sec, compared to 5 sec of Kinetics.

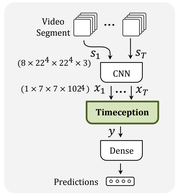
EXTENT

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

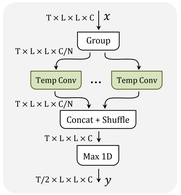
DEPENDENCY

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

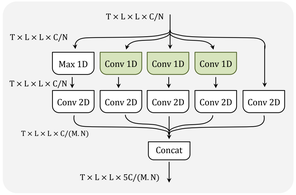
METHOD



(a) Model Overview



(b) Timeception Layer



(c) Temporal Conv Module

TEMPORAL-ONLY CONV

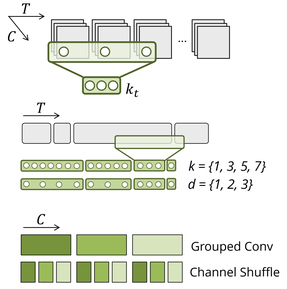
Depthwise separable 1D conv to reduce complexity of 3D conv from $O(t \cdot c^2)$ to $O(t \cdot c)$.

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.



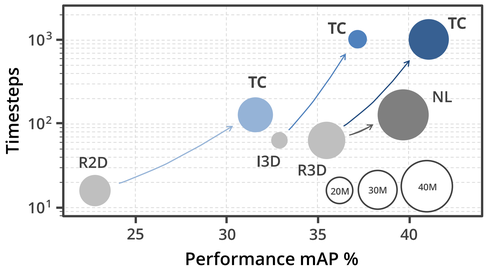
RESULTS

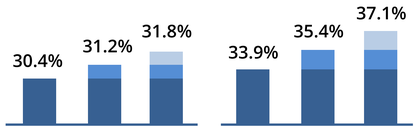
DATASET: CHARADES

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.





ResNet + Timeception

I3D + Timeception

LAYER EFFECTIVENES

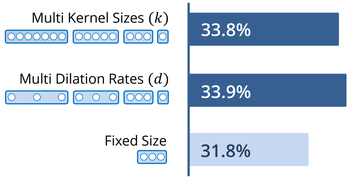
Timeception monotonically improves as the network goes deeper.

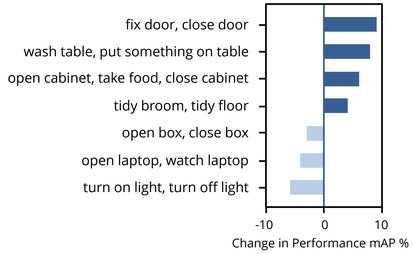
The same result is confirmed when using different backbones, as ResNet and I3D.

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.