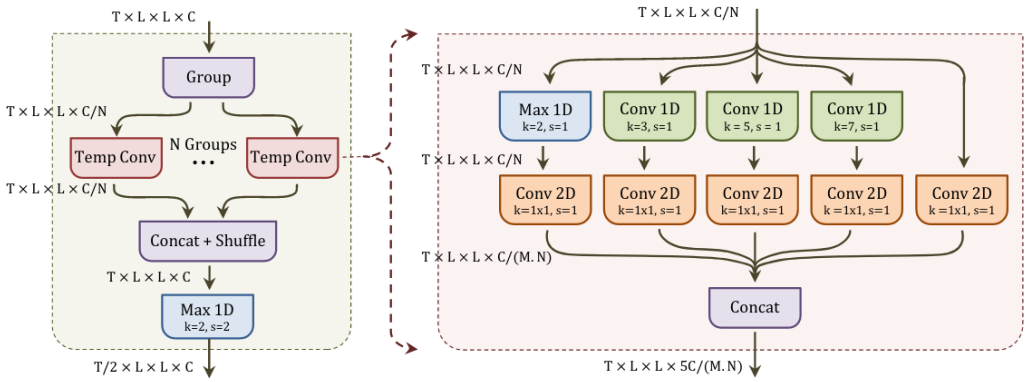
# Timeception for Complex Action Recognition

Noureldien Hussein, Efstratios Gavves, Arnold W.M. Smeulders

The paper proposes a new Timeception layer that can be added on top of standard video recognition backbones to process long-range temporal information efficiently for long-range action recognition. More specifically, the paper focuses on complex action recognition, which is defined by three main properties:

- composition: a single instance or video consists of several actions (get -> cook -> put -> wash).
- order: a given action can't be recognized by only focusing on the previous one, but it needs a larger temporal extent and dependency to detect.
- extent: actions vary in their temporal extents from one example to the other.




(a) Timeception Layer (b) Temporal Conv Module

First, to design such a model, the paper follows three key design principles:

- *Subspace Modularity:* In the context of deep network cascades, a decomposition should be modular, such that after a cascade of spatial and temporal convolutions, it must be possible that yet another cascade (of spatial and temporal convolutions) is possible and meaningful.
- *Subspace Balance:* Increasing the number of parameters for modeling a specific subspace should come at the expense of reducing the number of parameters of another subspace. For example, with 2D CNN, spatial subspace (S) is reduced while the semantic channel subspace (C) is expanded.
- *Subspace Efficiency:* the majority of the available parameter budget should be dedicated to subspaces that are directly relevant to the task at hand. For the task of long-range temporal modeling, the temporal subspace is the most important one.

Based on these design choices, the authors propose a **Timeception** layer which, 1) only processes the relevant subspace by solely using depthwise-separable temporal convolutions (called temporal convolutions in the paper) with a kernel T x 1 x 1 x 1 with no spatial kernels. 2) with the usage of successive temporal convolutions, there needs to be some subspace modularity since the semantic

subspace is ignored. To solve this, the Timeception module consists of a channel-wise dimensionality reduction which is then followed by the temporal convolutions, and to maintain efficiency, both are only applied over groups of channels, the output is then shuffled to mix the channels between groups before feeding the outputs to the Timeception second layer. 3) to take into consideration different temporal extents, the Timeception layer consists of parallel branches where each branch has a different kernel or dilation size to have a variable temporal receptive field, the outputs are then concatenated per group, and then all the groups are concatenated and shuffled.