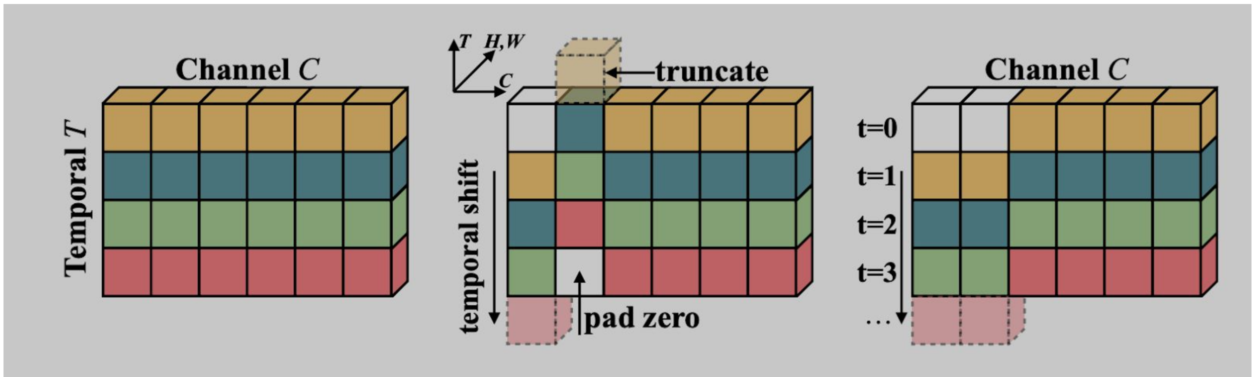
TSM: Temporal Shift Module for Efficient Video Understanding

Ji Lin | Chuang Gan | Song Han
(A paper review by Gautam Gare)



(a) The original tensor without shift. **(b)** Offline temporal shift (bi-direction). **(c)** Online temporal shift (uni-direction).

Action recognition in videos is the current topic of interest among computer vision researchers and a natural logical progression after tackling the problem of object recognition in images. Videos are a form of 3D data (time x height x width) source which conventionally are analyzed using 3D convolutional (conv) neural networks (CNNs). But such 3D conv blocks are slower and computationally heavier, thus needing GPUs with higher FLOPS and memory. In order to address this problem, Ji Lin et al. came up with TSM module, a novel solution to employ 2D convs to analyze 3D video data, which is faster and efficient while achieving better accuracy than using 3D CNNs for recognizing actions in videos.

TSM modules operate on the video, by processing each time frame independently using 2D conv blocks. The temporal information is infused into the CNN by intermixing the 2D features calculated for each time frame, with each other at every conv block in the CNN. For time step t, we introduce past info by replacing certain propositions of features with features corresponding to time step t-1. Similarly, we introduce future info by replacing certain propositions of features with features corresponding to time step t+1. For online video processing, only the past info is mixed, termed as uni-direction. While for offline video processing, both the past and future info is mixed, termed as bi-direction, as seen in the above figure. Their approach of independently calculating 2D features and intermixing temporal information beat the then state-of-the-art 3D CNNs by achieving 52.6 mAP on the Something-Something video dataset. Making real-time video action recognition a reality using CNNs while still remaining efficient and cost-effective.

Reference: Lin, Ji, Chuang Gan, and Song Han. "TSM: Temporal Shift Module for Efficient Video Understanding." *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019): 7082-7092.