TSM: Temporal Shift Module for Efficient Video Understanding - Summary

Ji Lin, Chuang Gan, Song Han

Introduction

The main problem pointed in this work is the need to model the temporal axis efficiently in order to achieve a better understanding in videos. In order to approach this problem, the authors run a quick comparison between 2D CNNs and their 3D counterpart. As a very short summary, the first type of CNNs are more efficient, but they are not able to model the temporal dimension well, especially when they are performing on single frames as input. The second type of model is more suited to learn spatial and temporal features, but the computation cost is way larger, making the deployment of such models almost impossible in real-world scenarios.

This paper approaches these two issues previously described by proposing a novel perspective for temporal modelling called the Temporal Shift Module (TSM). Specifically, a traditional 2D CNN operates independently over the time dimension, wile the TSM is used to shift the channels along this dimension, both in the past and in the feature. This results in an exchange of temporal information between consecutive frames, allowing the CNN model to learn temporal cues along with spatial information. The authors explain their intuition by breaking the convolution operation in two separate operations: shifting and multiply-accumulation. They argue that they shift the time dimension by 1 and fold the multiply-accumulate step from time dimension to channel dimension.

Temporal Shift Module

Firstly, the paper explains that naive shift of channels is not working. The main reasons are that:

- a simple naive shift is worsening the efficiency of the model due to large data movement between frames.
- although the temporal modelling capacity is increased, the performance is decreasing due to work spatial modelling ability. This is caused by the loss of spatial information involved in the shift process.

To tackle these two problems, two technical contributions are discussed next:

- **Reducing Data Movement:** In order to study the effect of data movement, the authors shifted different proportions of the channels and measured the resulted latency. As a result, they observed that if they shift only a small proportion of the channels (1/8), the latency overhead is increased the latency overhead with only 3%, which motivates the use of partial shift.
- **Keeping Spatial Feature Learning Capacity:** As pointed previously, there is a need to balance the model capacity for spatial feature learning and temporal feature learning. In order to do this, the TSM module is inserted inside the residual brach in a residual block, instead of being inserted before each convolutional layer. This type of residual shift addresses the degraded spatial feature learning problem, as all the information in the original activation is still accessible after temporal shift through identity mapping.

Offline Models with Bi-directional TSM

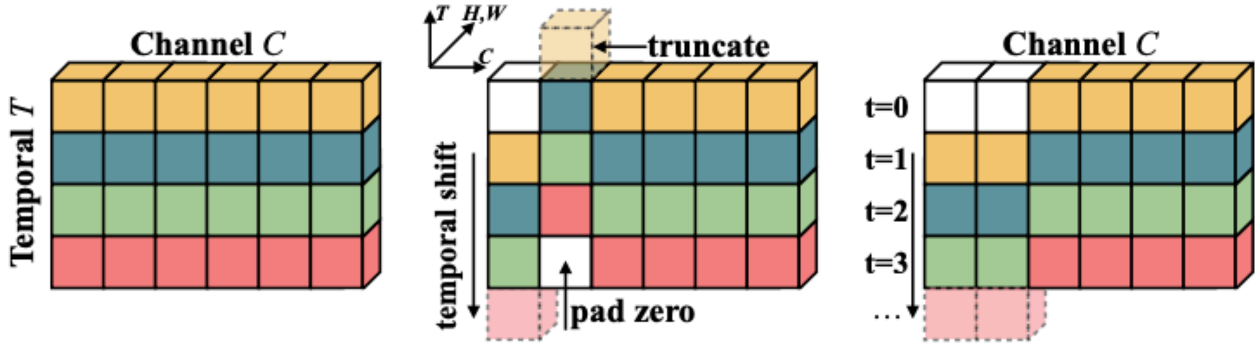
Given a video V, following the TSN method, T frames are sampled from the video. After sampling, the 2D CNN baseline processes each of these frames individually, and the output logins are averaged to give the final prediction. During the inference of the convolutional layers, the frames are still running independently, but a TSM module is inserted for each residual block, which enables temporal information fusion, at no computation. For each inserted temporal shift module, the temporal receptive field is enlarged by 2, as if running a convolution with the kernel size of 3 along the temporal dimension.

Online Models with Uni-directional TSM

As described previously, the offline version shifts channels bi-directionally, which requires features from future frames to replace the features in the current frame, which is not possible in a real-time scenario. In order to solve this issue, the shift will take place in only one direction, thus the model will use only the past in order to understand the action.

More specifically, during inference, for each frame, the first 1/8 feature maps of each residual block is saved and coached in the memory. For the next frame, 1/8 of the current feature maps is replaced with the cached features.

Temporal Shift Module:



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

Figure 1. **Temporal Shift Module (TSM)** performs efficient temporal modeling by moving the feature map along the temporal dimension. It is computationally free on top of a 2D convolution, but achieves strong temporal modeling ability. TSM efficiently supports both **offline** and **online** video recognition. Bi-directional TSM mingles both past and future frames with the current frame, which is suitable for high-throughput offline video recognition. Uni-directional TSM mingles only the past frame with the current frame, which is suitable for low-latency online video recognition.