D-STEP: Dynamic Spatio-Temporal Pruning

Avraham Raviv, Yonatan Dinai, Igor Drozdov, Niv Zehngut and Ishay Goldin

SAMSUNG

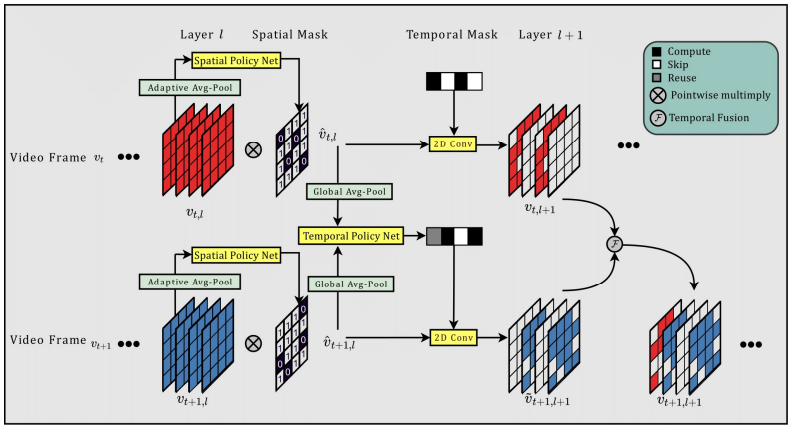
SIRC – **S**amsung **I**srael **R&D** **C**enter, Tel Aviv, Israel

BMVC 2022

MOTIVATION

- The exponential growth in online **video applications** creates a huge demand for systems that can recognize and localize actions, events, objects, and interactions in video.
- Many work handle those heavy tasks by using Transformers or 3D conv, but devices with limited resources, such as mobile phones, cameras and AR/VR glasses, **require more efficient algorithms**.
- By combining **spatial** and **temporal** modelling together, a neural network can gain a better understanding of the scene with no increase in computation.
- Spatio-temporal modeling can also be used to **identify redundant and sparse information** in both the spatial and the temporal domains.

METHOD

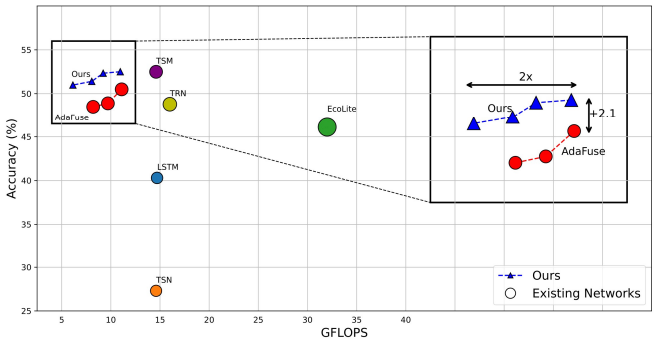


Using two policies networks, we prune spatial and temporal redundancy and aggregate information during frames while paying attention to spatial maps S. Each feature map at layer ℓ is first filtered through S. Then, at time $t+1$, the 2D Conv layer processes compute channels (blue) in feature map $\mathcal{V}_{t+1,\ell}$, and fuses the reuse channels (red) from the history feature map $\mathcal{V}_{t,\ell+1}$.

OBJECTIVES

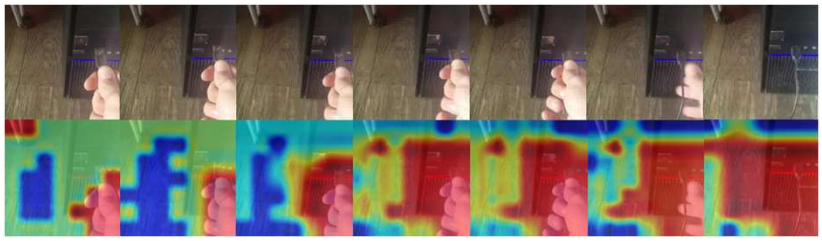
- **Dynamically identify spatial and channel sparsities as well as temporal redundancies in order to prune tiles and filters, minimizing computations.**
- **Utilize the concept of temporal aggregation as a method of improving accuracy.**
- **Develop a generic approach for all spatio-temporal tasks, making it suitable for a wide range of video-specific architectures.**
- **improved accuracy-compute trade-off over the current state-of-the-art methods.**

COMPARISON WITH SOTA METHODS



Something-V2 Validation Accuracy vs FLOPs.

QUALITIVE RESULTS – SPATIAL SPARSITY



The top row shows a random clip, while the bottom row shows the averaged Gumbel Sigmoid outputs across all layers of the network. Regions marked in red correspond to areas with high computation while blue regions were mostly skipped.

Visit Us:

ABLATION STUDY

Model	#Params	FLOPs	Top 1	Top5
Baseline – Static Model [TSN]	11.2M	14.6G	27.3	38.0
Static TS + DCP	15.6M	10.78G	51.96	80.12
Dynamic TS + DCP [AdaFuse]	15.6M	11.1G	50.5	67.8
Dynamic TS + DCP + SS	15.6M	9.32G	51.29	79.15
Static and Dynamic TS + DCP	15.6M	11.02G	50.92	78.9
Static and Dynamic TS + DCP + SS	15.6M	**9.22G**	**52.34**	**80.48**

Ablation study of dynamic inference components using ResNet18 on Something-V2 dataset. DCP - Dynamic Channel Pruning. TS - Temporal Shift. SS - Spatial Sparsity.