Skip-Convolutions for Efficient Video Processing - Supplementary Material

Amirhossein Habibian Davide Abati Taco S. Cohen Babak Ehteshami Bejnordi

Qualcomm AI Research*

{habibian,dabati,tacos,behtesha}@qti.qualcomm.com

1. Experiments on video classification

Although our work is mainly focused on stream processing, *i.e*. video tasks where a spatially dense prediction is required for every frame, our Skip-Conv model can in principle enable improvement in efficiency of classification models. We hereby conduct a preliminary experiment with video classification on human action dataset [2], and consider the split-1 using RGB modality. We report Top-1 accuracy for center-crop inference. As a backbone model, we rely on Temporal Segment Networks (TSN) [6], based on a ResNet-101. We study the performance of Skip-Conv in two inference setup: *i.* TSN-25, where inference is carried out over 25 frames per clip (sampled uniformly from the whole video as in [6]). *ii.* TSN-6, where inference is carried out over 6 frames per clip, so there are much less redundancies between frames.

As reported in Tab. 1, Skip-Conv reduces the computation cost with a minor accuracy drop. The computation gain is higher for the high frame-rate model (TSN-25) as there are more redundancies between video frames. For the low frame-rate model (TSN-6), the compute is reduced from 46.80 to 35.83 GMACs even though the frame redundancies and residual sparsities are fairly low because of the coarse frame sampling.

Finally, we remark that many state-of-the-art video classification architectures rely on 3D backbones [1, 3, 4]. Although the focus of this paper has been on 2D Skip-Convs, Eq. 1 and 2 in the main paper can be extended to the case of 3D convolutions as a linear operator.

	Accuracy	**GMAC**
TSN-25	55.49	194.98
TSN-25 + Skip-Conv	54.77	41.02
TSN-6	53.84	46.80
TSN-6 + Skip-Conv	53.77	35.83

Table 1: Video classification results. Skip-Conv reduces the computation cost with a minor accuracy drop.

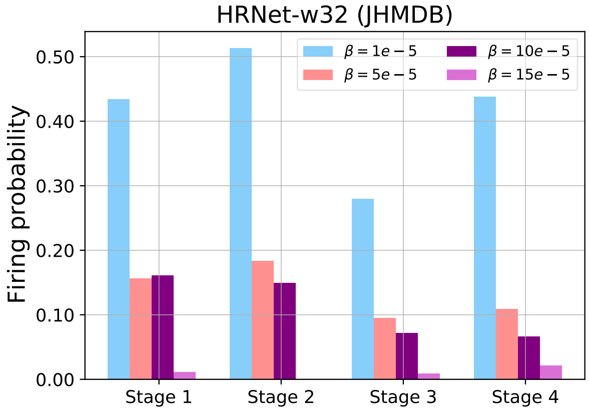

Figure 1: Per-stage sparsity level of HRNet-32.

2. Sparsity ratios

In this section we analyze the amount of sparsity induced by Skip-Conv in different levels of a backbone network. To this end, we refer to the pose estimation experiments described in Sec. 4.2 of the main paper, and we rely on the same setting by considering the JHMDB dataset [2] with a HRNet-w32 backbone network [5]. We train the Skip-Conv model with Gumbel gates under different sparsity objectives, by varying β in $[1e-5, 5e-5, 10e-5, 15e-5]$. For completeness, we also report the performance of these models, that score $[0.95, 0.94, 0.93, 0.91]$ in PCK respectively. We then measure how the firing probability of gates in Skip-Conv changes at different depths in the network.

The results are summarized in Fig. 1, where we report the probability of firing at different stages of the base HRNet-w32 model, averaged over all test examples, different layers within the same stage, and the three splits commonly used in pose estimation protocols. The figure highlights how, in general, Skip-Conv allows to bypass a significant amount of computation. Even under very mild sparsity constraints (*i.e.* $\beta = 1e-5$), the Gumbel gates learn to skip more than half of the pixels in feature maps overall. For intermediate values of β, firing probabilities drop to below 0.2, and in some cases fall under 0.1 for later stages in the network. For high

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

sparsity coefficients (*i.e.* $\beta = 15e - 5$) Skip-Conv mostly relies on features from the first frame in the input clip, and triggers computation very occasionally (0.8% to 2.1% of pixels). Interestingly, for stage 2 has a firing probability of zero: this means that the model only relies on features from the reference frame for those layers, and that they suffice to carry out correct predictions.

References

[1] C. Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 1

[2] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 1

[3] A. Piergiovanni, A. Angelova, and M. S. Ryoo. Tiny video networks. *arXiv preprint arXiv:1910.06961*, 2019. 1

[4] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 1

[5] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 1

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1