Conditional Model Selection for Efficient Video Understanding: Supplementary material

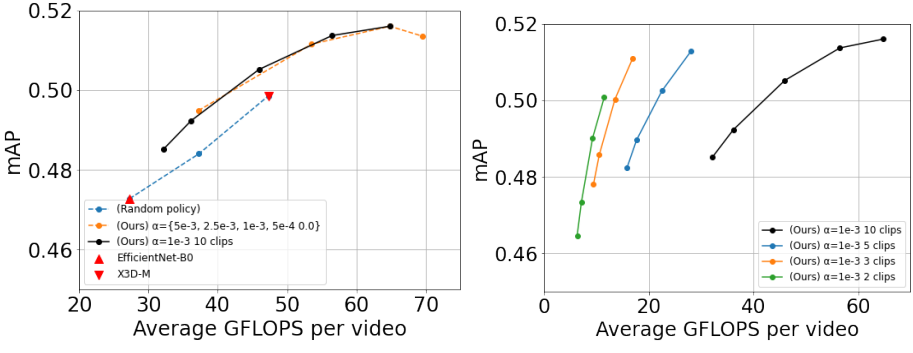
Mihir Jain[1]*
Haitam Ben Yahia[1]*
Amir Ghodrati[1]
Fatih Porikli[2]
Amirhossein Habibian[1]

[1] Qualcomm AI Research[§],
Qualcomm Technologies Netherlands B.V.

[2] Qualcomm AI Research[§],
Qualcomm Technologies, Inc.

{mijain,hyahia,ghodrati,fporikli,ahabibia}@qti.qualcomm.com

In this supplementary material, we first present further implementation details for classification and ablations on HVU dataset. Finally, we show some qualitative results for localization.



(a) (b)

Figure 1: **HVU ablations**: In (a) we show an ablation of α. In (b) we have an ablation on the number of clips aggregated in HVU.

1 Additional implementation details

Classification: As mentioned in Section 4.1 of the paper, we use pre-trained models on HVU, these are trained as follows. First, we extract features from the first classification layer after convolutions, from both models. A final classification layer is then added in the form of a $3 \times t$ convolution over the features dimension followed by an MLP, where t equals the temporal dimension of the feature, i.e. 7 (number of frames) in the case of EfficientNet-B0

* Equal Contribution. Correspondence to hyahia@qti.qualcomm.com

§ Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

and 1 (clip) for X3D. Since HVU is multi label based, we use the max function to aggregate clips. This results in the final prediction for a model in the model pool. For Kinetics we don't need to finetune the models as they are pretrained on Kinetics and our clips are aggregated by averaging over them.

Localization: For localization during inference, we follow the two-stage thresholding scheme of [1]. The first threshold is applied to filter out the classes that have video-level scores less than the average over all the classes. The second threshold is applied along the temporal axis to obtain the start and the end of each action instance.

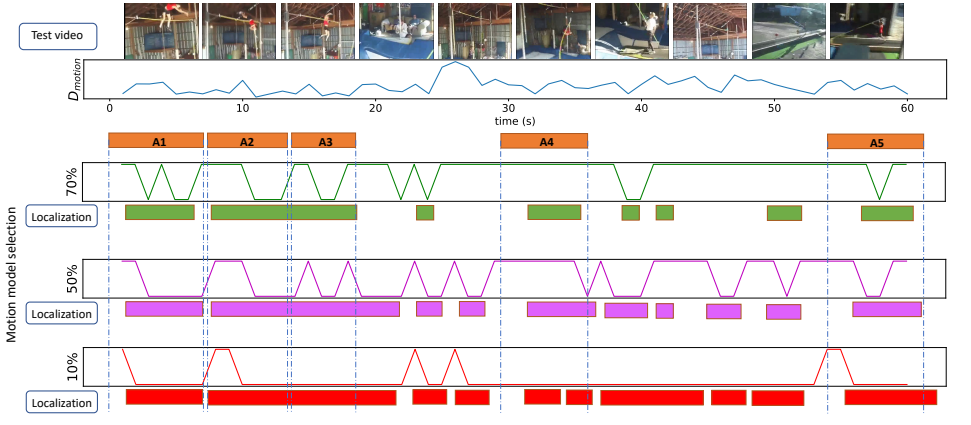


Figure 2: **Localization with varying use of motion model:** Top row shows first 60 seconds of an example video containing five instances of action *PoleVault*, A1 to A5 shown by orange boxes. Second row shows disparity, D_{motion}, as a function of time. Then, motion model selection is shown by applying varying thresholds on the confidence scores of model-selector. Thus, output of model-selector is shown for when the ratio of motion model selection is 70%, 50% and 10%. *For 10% case*, only instances A1 and A5 are detected for $IoU = 0.5$, while other three are false-negatives along with eight false-positives. *For 50% case*, A1, A4 and A5 are localized, along with seven false-positives. *Finally for 70% case*, A2 is also detected in addition and only A3 is missed. Number of of false-positives are also decreased to four.

2 Ablations

HVU ablations: In Figure 1-a we first show that our model outperforms the Random selection baseline as well as the models that are used in the model pool. Further more, when we set α in the range $5e^{-3} - 5e^{-5}$ we see that the performance roughly corresponds to the results of $\alpha = 5e^{-4}$ at test time. This also suggests, using a reasonable α, we can get similar results close to this α at test time without having to retrain. In Figure 1-b we vary the number of clips. Here we also see that our method has evidence of being complimentary to methods that act on salient frames or clips. All of these conclusions are also in line with the results for Kinetics shown in Section 4.2 in the paper.

3 Qualitative results

In Figure 2, on top we show an example input frame sequence, motion disparity values (D_{motion}) over time and temporal ground-truth of action instances. Then, motion model selections and the localization results are shown for varying thresholds on the confidence scores of model-selector. These are shown for when the ratio of motion model selection is 70% (green), 50% (magenta) and 10% (red). Evaluating at $IoU = 0.5$, only instances A1 and A5 are localized with 10% motion model selection, costing 8 false-positives. With 50% motion model selection, instance A4 is localized additionally, thanks to the motion model selection near $t = 33s$. There are still 7 false-positives. Finally, with 70% motion model selection, all the instances except A3 are localized correctly. Number of false-positives is also decreased to 4. Here, compared to 50% case, A2 is localized due to model motion selection just after instance A3. Model-selector predictions more often than not match the motion disparity, i.e., motion model is predicted near peaks of D_{motion} and for its lowest values appearance model is predicted. Training the model-selector with this approximate supervision signal leads to effective selection of models for action localization.

References

[1] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. 2018.