Efficient Action Recognition via Dynamic Knowledge Propagation -Supplementary Material-

Hanul Kim[1,2], Mihir Jain[1], Jun-Tae Lee[1], Sungrack Yun[1], Fatih Porikli[1]
[1]Qualcomm AI Research*
[2]Seoul National University of Science and Technology
hukim@seoultech.ac.kr, {mijain, juntlee, sungrack, fporikli}@qti.qualcomm.com

In this supplementary material, we first present an ablation study on the impact of different frame sampling strategies on the proposed approach. Finally, we show some additional qualitative results.

S-1. Frame sampling strategies

We consider two strategies for frame sampling. First, as used in the main paper, we adapt sampling intervals r_s and r_t across videos in order to have the same n_s and n_t for all the videos. For this 'adaptive' strategy, we plot for $n_s/n_t = 4$ as *Adaptive-4* (used in the main paper) and $n_s/n_t = 8$ as *Adaptive-8*. The mAP-GFLOPs curves for our method are shown in Figure S-1. Second, we fix the sampling intervals r_s and r_t for all the videos, as a result n_s and n_t vary across videos and are proportional to the video-length. While n_s and n_t vary over videos, we set the ratio n_s/n_t as 4 and 8 to plot them as *Fixed-4* and *Fixed-8*, respectively.

As Figure S-1 illustrates, given enough sampled frames *i.e*. beyond 12 GFLOPs, all four plots of the two sampling strategies achieve similar and promising performances. However, Adaptive-4 and Adaptive-8 experience larger performance drop at lower GFLOPs. This is because, in this setup, only a few sampled frames $n_t = 3$ are available per video, which leaves longer videos under-sampled. On the contrary, the fixed sampling interval strategy alleviates this problem by adjusting the number of sampled frames n_s and n_t according to the video-length, and achieves better mAP over lower computation range. Also, we see that Adaptive-8 and Fixed-8 perform a bit better than Adaptive-4 and Fixed-4 at the low GFLOPs setting, respectively. This shows more sampled frames for student is better in the low computation range.

Figure S-2 analyzes the impact of the number of sampled frames n_s and n_t by plotting mAP-GFLOPs curves. Specifically, *nt-5* sets n_t as 5 and varies n_s as $\{5, 20, 35, 50\}$. Similarly, *ns-20* sets n_s as 20 and varies n_t as $\{5, 10, 15, 20\}$.

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

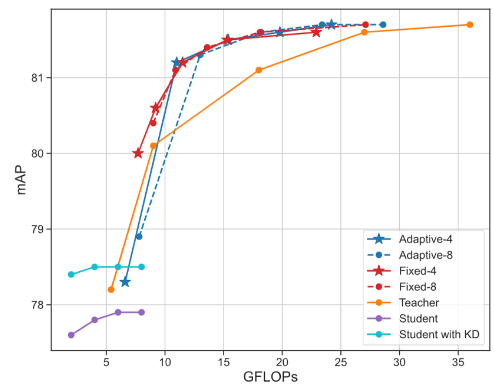


Figure S-1: Impact of Adaptive and Fixed sampling strategies (on ActivityNet 1.3): For both strategies, we set the ratio n_s/n_t to 4 and 8, and plot for accuracy vs. efficiency.

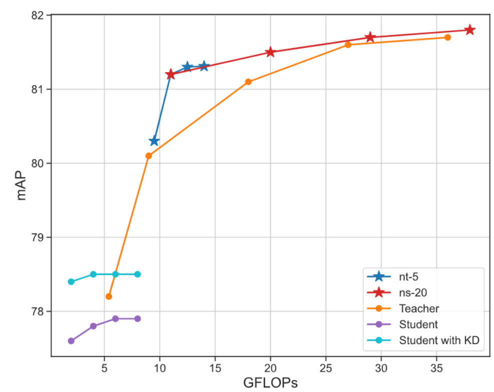
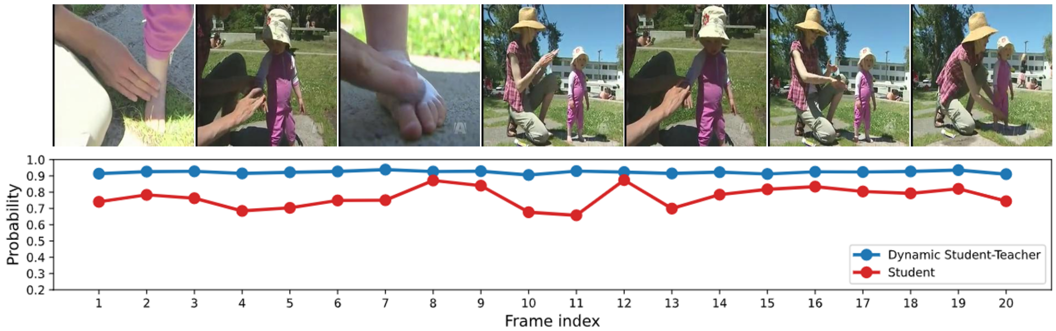
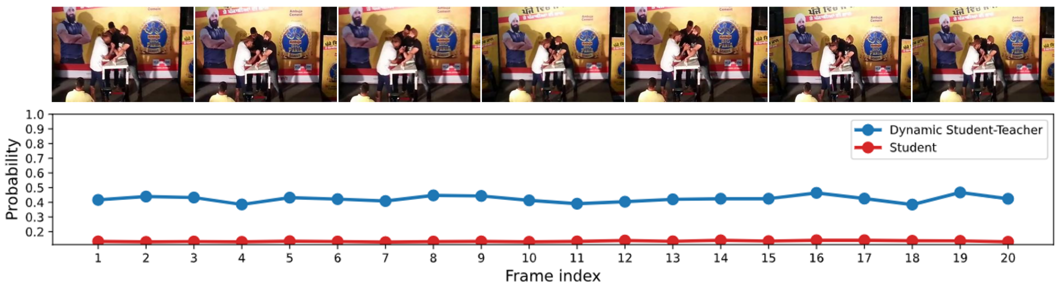


Figure S-2: Impact of number of samples frames on accuracy vs. efficiency plots on ActivityNet 1.3. We fix $n_t = 5$ and vary n_s from 5 to 50 for "nt-5". We set $n_s = 20$ and vary n_t from 5 to 20 for "ns-20". Naturally, we adopt the adaptive sampling intervals approach for both settings.



(a) Applying Sunscreen



(b) Arm wrestling

Figure S-3: Action recognition on two videos: (a) Applying Sunscreen and (b) Arm wrestling. In both examples, the first row illustrates input frames, and the second row shows the sequence of probabilities for ground-truth class predicted by Student and Dynamic Student-Teacher Ensemble.

Both these settings outperform Teacher at lower computation range, showing valued utilization of the sampled frames for student by the proposed dynamic knowledge propagation. Though n_t needs to to be increased to continue to improve mAP, as seen in case of *ns-20*, which goes on to better the Teacher even at higher GFLOPs.

S-2. Additional Qualitative Results

Figure S-3 shows qualitative results and the frame-level probabilities indicating that a frame belongs to the ground-truth action class. The proposed method provides more accurate frame-level predictions through dynamic knowledge propagation to convey the teacher's knowledge to the student during the inference time. Especially in Figure S-3 (b), the student network fails to yield accurate frame-level predictions. On the contrary, the proposed method provides relatively high probabilities for ground-truth class by exploiting more reliable information from the teacher network.