AdaFrame: Adaptive Frame Selection for Fast Video Recognition (Supplemental Material)

Zuxuan Wu[1*], Caiming Xiong[2†], Chih-Yao Ma[3], Richard Socher[2], Larry S. Davis[1]
[1] University of Maryland, [2] Salesforce Research, [3] Georgia Institute of Technology

Stop criterion using predicted utilities and entropy

As mentioned in the main paper, stopping inference based on entropy fails to enable adaptive inference. We visualize the fraction of sample (frequency) classified over time using both utility and entropy in Figure 1. For utility-based stopping, we can easily get a "bell curve" with different μ; similar behavior cannot be observed by varying threshold for entropy.

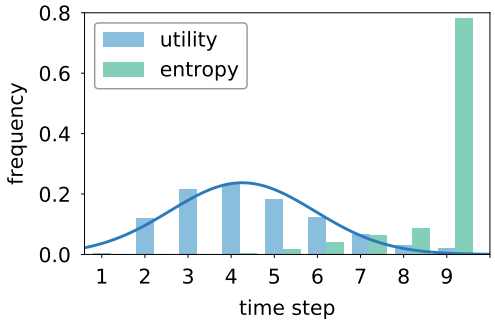

Figure 1: **Fraction of samples classified over time using utility and entropy.**

Detailed Results on FCVID and ActivityNet

	μ	mAP	GFLOPs	# Frames
AvgPooling	—	64.2	15.64 ± 0.00	2.00 ± 0.00
LSTM	—	64.2	15.68 ± 0.00	2.00 ± 0.00
AdaFrame-3	0.2	66.1	15.26 ± 5.05	1.67 ± 0.55
AdaFrame-3	0.5	72.1	17.64 ± 2.50	1.93 ± 0.27
AdaFrame-3	0.7	**76.5**	25.14 ± 5.02	2.75 ± 0.55
AvgPooling	—	76.3	39.10 ± 0.00	5.00 ± 0.00
LSTM	—	76.0	39.19 ± 0.00	5.00 ± 0.00
AdaFrame-5	0.2	74.9	24.13 ± 5.04	2.64 ± 0.55
AdaFrame-5	0.5	76.6	31.62 ± 6.78	3.46 ± 0.74
AdaFrame-5	0.7	**78.6**	44.96 ± 6.18	4.92 ± 0.68
AvgPooling	—	78.9	78.20 ± 0.00	10.00 ± 0.00
LSTM	—	78.1	78.40 ± 0.00	10.00 ± 0.00
AdaFrame-10	0.2	77.9	37.84 ± 9.56	4.14 ± 1.04
AdaFrame-10	0.5	78.8	45.33 ± 18.74	4.96 ± 2.05
AdaFrame-10	0.7	**80.2**	75.04 ± 16.45	8.21 ± 1.80

Table 1: **Detailed results on FCVID.**

	μ	mAP	GFLOPs	# Frames
AvgPooling	—	50.5	15.64 ± 0.00	2.00 ± 0.00
LSTM	—	53.0	15.66 ± 0.00	2.00 ± 0.00
AdaFrame-3	0.2	56.3	14.97 ± 4.37	1.64 ± 0.48
AdaFrame-3	0.5	61.2	17.80 ± 1.85	1.95 ± 0.20
AdaFrame-3	0.7	**64.1**	26.11 ± 3.54	2.86 ± 0.39
AvgPooling	—	66.1	39.09 ± 0.00	5.00 ± 0.00
LSTM	—	67.0	39.16 ± 0.00	5.00 ± 0.00
AdaFrame-5	0.2	66.0	24.37 ± 4.56	2.67 ± 0.50
AdaFrame-5	0.5	68.7	31.49 ± 6.32	3.45 ± 0.69
AdaFrame-5	0.7	**69.5**	34.69 ± 4.25	3.80 ± 0.47
AvgPooling	—	68.6	78.18 ± 0.00	10.00 ± 0.00
LSTM	—	70.4	78.32 ± 0.00	10.00 ± 0.00
AdaFrame-10	0.2	68.1	45.10 ± 4.29	4.94 ± 0.47
AdaFrame-10	0.5	69.1	53.50 ± 6.84	5.86 ± 0.75
AdaFrame-10	0.7	**71.5**	78.97 ± 5.20	8.65 ± 0.57

Table 2: **Detailed results on ActivityNet.**

Table 1 and Table 2 present the detailed results of Figure 3 in the main paper on FCVID and ActivityNet, respectively. In particular, we train AdaFrame with fixed K time steps to obtain different models, denoted as AdaFrame-K to accommodate different computational requirements during testing; and for each model we vary μ such that adaptive inference can be achieved within the same model.

Enlarged Figure 3

Figure 2 presents the enlarged version of Figure 3 in the main paper.

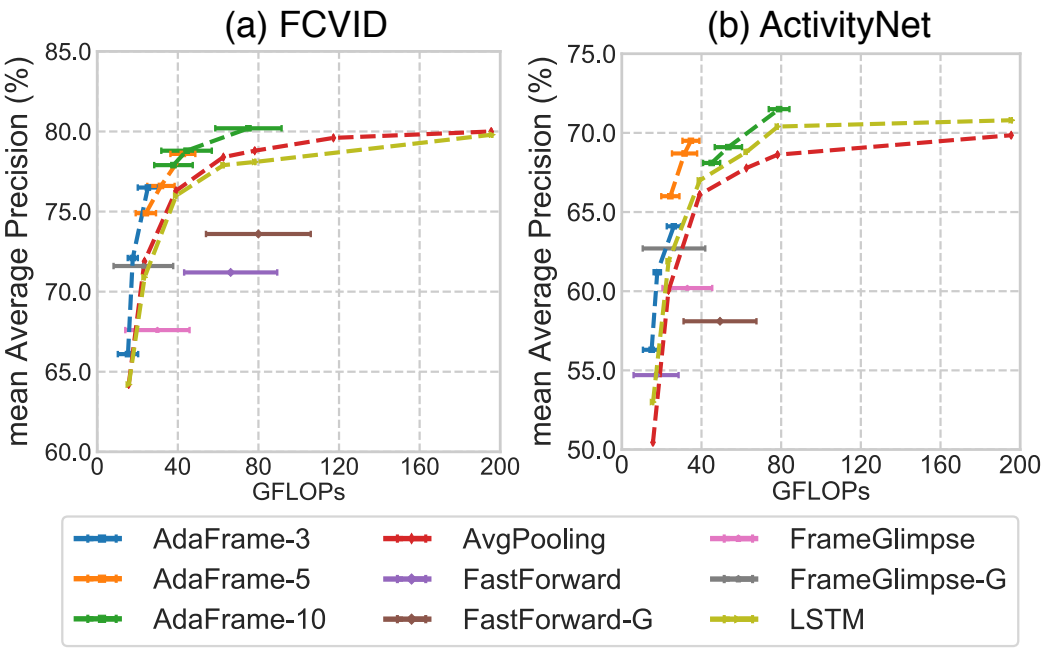

Figure 2: **Mean average precision *vs*. computational cost**. Comparisons of AdaFrame with FrameGlimpse, FastForward, and alternative frame selection methods based on heuristics.