Supplementary Material
FrameExit: Conditional Early Exiting for Efficient Video Recognition

Amir Ghodrati* Babak Ehteshami Bejnordi* Amirhossein Habibian
Qualcomm AI Research†
{ghodrati, behtesha, ahabibia}@qti.qualcomm.com

1. Algorithmic description of the policy

In this section, we present an algorithmic description of our policy function. Our policy function follows a coarse-to-fine principle for sampling frames. It starts sampling from a coarse temporal scale and gradually samples more frames to add finer details to the temporal structure. Specifically, we sample the first frames from the middle, beginning, and end of the video, respectively, and then repeatedly sample frames from the halves as shown in Algorithm 1.

Algorithm 1: Policy function

Input: N (number of frames in the video)
Output: **S** (a sequence of sampled frame indices)
S = [floor((N+1)/2), 1, N] ;
q = 2 ;
while *len(**S**) < N* **do**
  interval = floor(linspace(1, N, q+1));
  for *i=1 : len(interval)-1* **do**
    a = interval[i];
    b = interval[i+1];
    ind = floor((a+b)/2);
    if *ind is not in **S*** **then**
      S.append(ind);
    end
  end
  q = q * 2;
end

2. Wall-clock time inference

The wall-clock speedup of FrameExit is highly correlated with FLOPS. This is because the major computational cost comes from the number of frames being processed by the backbone. For example, using an Nvidia GeForce GTX 1080Ti GPU, the inference time of FrameExit with a ResNet50 backbone at computational costs of 41.2, 26.1, and 12.3 GFLOPs are 15.8, 10.9, and 4.8 ms, respectively.

3. Qualitative results

Figure 1 presents several video clips and the number of frames required to process by FrameExit before exiting. As shown, a few frames are sufficient for FrameExit to classify a video if the observed frames are discriminative enough. As the complexity of video scene increases (such as obscured objects and cluttered background) FrameExit needs more temporal information to reach a confident prediction.

Figure 1: **Qualitative results.** Each row contains 10 frames of a video, sampled according to the policy function. Our method observes only the green box to recognize an action. As the content becomes more complex, FrameExit needs more frames to make a confident prediction. Zoom in for higher quality. Videos are adopted from [1, 2, 3, 4, 5].

References

[1] ORBEA ALMA M30 CARBON · 2014 / BICIMAG by bicimag is licensed under CC BY. 1
[2] Sophie and Greg playing acro yoga by Gregology is licensed under CC BY. 1
[3] Kilby Plays the Harmonica - Sort of by aplusjimages is licensed under CC BY. 1
[4] Batlle of The Year: Dau Truong Breakdance TV Spot by Galaxy Studio is licensed under CC BY. 1
[5] Show de tango COOPRUDEA by Cooperativa de Profesores Cooprudea is licensed under CC BY. 1

*Equal contribution
†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc