Appendix for “Adaptive Focus for Efficient Video Recognition”

A. Introduction of Baselines

AdaFocus is compared with several competitive baselines that focus on facilitating efficient video recognition, including MultiAgent [9], SCSampler [4], LiteEval [11], AdaFrame [10], Listen-to-look [1] and AR-Net [6].

- MultiAgent [9] proposes to learn to select important frames with multi-agent reinforcement learning.
- SCSampler [4] introduces a light-weighted framework to efficiently identify the most salient temporal clips within a long video. We follow the implementation of [6].
- LiteEval [11] combines a coarse LSTM and a fine LSTM to adaptively allocate computation based on the importance of frames.
- Listen-to-look [1] fuses image and audio information to select the key clips within a video. As we do not leverage the audio of videos, for a fair comparison, we adopt its image-based version introduced in their paper.
- AR-Net [6] dynamically identifies the importance of video frames, and processes them with different resolutions accordingly.

B. Implementation Details

B.1. Training Hyper-parameters for Section 4.1

In our implementation, we always train $f_G$, $f_L$ and $f_C$ using a SGD optimizer with cosine learning rate annealing and a Nesterov momentum of 0.9. The size of the mini-batch is set to 64, while the L2 regularization coefficient is set to 1e-4. We initialize $f_G$ and $f_L$ by fine-tuning the ImageNet pre-trained MobileNet-V2 [7] and ResNet-50 [2] using full inputs for 15 epochs with an initial learning rate of 0.01. In stage I, we train $f_L$ and $f_C$ using randomly sampled patches for 50 epochs with an initial learning rate of 5e-4 and 0.05, respectively. Here we do not train $f_G$ as we find this does not significantly improve the performance, but increases the training time. In stage II, we train $\pi/\pi'$ with an Adam optimizer [3] for 50/10 epochs. The same training hyper-parameters as [8] are adopted. In stage III, we only fine-tune $f_C$ with the learned policy for 10 epochs, since we find further fine-tuning $f_L$ leads to trivial improvements but prolongs the training time. The initial learning rates are set to 5e-4 and 5e-3 for Mini-Kinetics and ActivityNet/FCVID, respectively.

B.2. Training Hyper-parameters for Section 4.2

Here we initialize $f_G$ and $f_L$ by training them using the same configuration as [5]. The training procedure of AdaFocus is the same as Section 4.1 except for the following changes. In stage I, we use the initial learning rate of 1e-5 and 0.01 for $f_L$ and $f_C$, respectively, and train them for 10 epochs. In stage III, we use an initial learning rate of 5e-4 for $f_C$. Note that TSM+ follows exactly the same training procedure as our method. The only difference is that TSM+ does not train the policy network $\pi$, since it adopts full frames as inputs.

References


