Appendices for "AdaFocusV3: On Unified Spatial-temporal Dynamic Video Recognition"

Yulin Wang[1⋆], Yang Yue[1⋆], Xinhong Xu[1], Ali Hassani[2], Victor Kulikov[3], Nikita Orlov[3], Shiji Song[1], Humphrey Shi[2,3†], and Gao Huang[1,4†]

[1] Department of Automation, BNRist, Tsinghua University
[2] University of Oregon
[3] Picsart AI Research (PAIR)
[4] Beijing Academy of Artificial Intelligence (BAAI)
{wang-yl19, ley18}@mails.tsinghua.edu.cn, shihonghui3@gmail.com, gaohuang@tsinghua.edu.cn

A Experimental Setups

Datasets. Our experiments are based on six large-scale video recognition benchmark datasets, *i.e.*, ActivityNet [1], FCVID [4], Mini-Kinetics [5,15], Something-Something (Sth-Sth) V1&V2 [3] and Diving48 [7]. The official training-validation split is adopted for all of them. Note that these datasets are widely used in the experiments of a considerable number of recently proposed baselines. We select them for a reasonable comparison with current state-of-the-art results.

– ActivityNet [1] contains the videos of 200 human action categories. It includes 10,024 training videos and 4,926 validation videos. The average duration is 117 seconds.
– FCVID [4] includes 45,611 training videos and validation 45,612 videos. The data is annotated into 239 classes. The average duration is 167 seconds.
– Mini-Kinetics is a subset of the Kinetics [5] dataset. It contains include 200 randomly selected classes of videos, with 121k videos for training and 10k videos for validation. The average duration is around 10 seconds [5]. We establish it following [15,10,11,9].
– Something-Something (Sth-Sth) V1&V2 [3] datasets contain 98k and 194k videos respectively. Both of them are labeled with 174 human action categories. The average duration is 4.03 seconds.
– Diving48 [7] is a fine-grained video dataset of competitive diving, consisting of ∼18k trimmed video clips of 48 unambiguous dive sequences.

Data pre-processing. We uniformly sample 18 frames from each video on ActivityNet, FCVID and Mini-Kinetics, while sampling 8/16 frames on Sth-Sth and Diving48. These configurations are determined on the validation set for a favorable accuracy-efficiency trade-off. The data augmentation pipeline in [8,10,11,12,13] is adopted. Specifically, the frames of training data is randomly scaled and cropped into 224×224 images. On all the datasets except for Sth-Sth V1&V2 and Diving48, the random flipping is performed as well. At test time, all the frames are resized to 256×256 and centre-cropped to 224×224.

⋆ Equal contributions. † Corresponding Authors.

B Baselines

Baselines. We compare AdaFocusV3 with a variety of recently proposed approaches that focus on improving the efficiency of video recognition. The results on ActivityNet, FCVID and Mini-Kinetics are provided. In addition to the previous versions of AdaFocus, the following baselines are included.

- LiteEval [15] dynamically activates coarse and fine LSTM networks conditioned on the importance of each frame.
- SCSampler [6] is an efficient framework to select salient video clips or frames. The implementation in [10] is adopted.
- ListenToLook [2] searches for the task-relevant video frames by leveraging audio information. We adopt the image-based variant introduced in their paper for fair comparisons, since we do not use the audio of videos.
- AR-Net [10] processes the frames with different resolutions based on their relative importance.
- AdaFrame [14] adaptively identifies the informative frames from the videos with reinforcement learning.
- VideoIQ [11] learns to process each frame with different precision based the importance in terms of video recognition.
- OCSampler [9] is a one-stage framework that learns to represent the video with several informative frames with reinforcement learning.

C Training Details

On ActivityNet, FCVID and Mini-Kinetics, the training of AdaFocusV3 exactly follows the same end-to-end training pipeline as AdaFocusV2 [13]. On Something-Something (Sth-Sth) V1&V2 and Diving48, we first train the two deep encoders and the classifier with a random policy, and then train the policy network isolatedly. We find that this two-stage training pipeline yields a better performance, while its training cost is approximately the same as its end-to-end training counterpart.

D More Results

Effectiveness of the early-termination algorithm is validated in Figure 1. The results on ActivityNet with the cube size of 128×128×1 are presented. Three variants are considered: (1) adaptive early-exit with prediction confidence; (2) random early-exit with the same exit proportion as AdaFocusV3; (3) early-exit with fixed cube number. Our entropy-based mechanism shows the best performance.

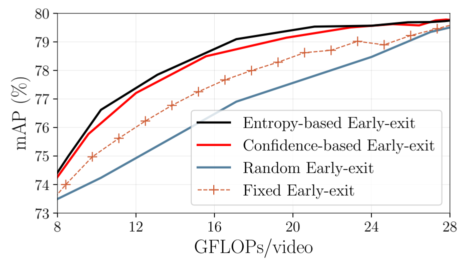

Fig. 1. Ablation study on early-termination algorithm.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970 (2015) 1
2. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: CVPR. pp. 10457–10467 (2020) 2
3. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: ICCV. pp. 5842–5850 (2017) 1
4. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 352–364 (2018) 1
5. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 1
6. Korbar, B., Tran, D., Torresani, L.: Scsampler: Sampling salient clips from video for efficient action recognition. In: ICCV. pp. 6232–6242 (2019) 2
7. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: ECCV. pp. 513–528 (2018) 1
8. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV. pp. 7083–7093 (2019) 1
9. Lin, J., Duan, H., Chen, K., Lin, D., Wang, L.: Ocsampler: Compressing videos to one clip with single-step sampling. In: CVPR (2022) 1, 2
10. Meng, Y., Lin, C.C., Panda, R., Sattigeri, P., Karlinsky, L., Oliva, A., Saenko, K., Feris, R.: Ar-net: Adaptive frame resolution for efficient action recognition. In: ECCV. pp. 86–104. Springer (2020) 1, 2
11. Sun, X., Panda, R., Chen, C.F.R., Oliva, A., Feris, R., Saenko, K.: Dynamic network quantization for efficient video inference. In: ICCV. pp. 7375–7385 (2021) 1, 2
12. Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., Huang, G.: Adaptive focus for efficient video recognition. In: ICCV (October 2021) 1
13. Wang, Y., Yue, Y., Lin, Y., Jiang, H., Lai, Z., Kulikov, V., Orlov, N., Shi, H., Huang, G.: Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. In: CVPR (2022) 1, 2

14. Wu, Z., Li, H., Xiong, C., Jiang, Y.G., Davis, L.S.: A dynamic frame selection framework for fast video recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020b) 2
15. Wu, Z., Xiong, C., Jiang, Y.G., Davis, L.S.: Liteeval: A coarse-to-fine framework for resource efficient video recognition. In: NeurIPS (2019b) 1, 2