Dynamic Network Quantization for Efficient Video Inference (Supplementary Material)

Ximeng Sun[1] Rameswar Panda[2] Chun-Fu (Richard) Chen[2]
Aude Oliva[2,3] Rogerio Feris[2] Kate Saenko[1,2]
[1]Boston University, [2]MIT-IBM Watson AI Lab, [3]MIT
Project page: https://cs-people.bu.edu/sunxm/VideoIQ/project.html

Section	Content
A	Dataset Details
B	Implementation Details
C	Additional Ablation Studies
D	Qualitative Results

Table 1: **Supplementary Material Overview**

Arch.	α_{init}	32-bit		4-bit		2-bit	
		α_{lr}	α_{wd}	α_{lr}	α_{wd}	α_{lr}	α_{wd}
ResNet-18	4	0.01	5e-4	0.01	5e-4	0.01	5e-3
ResNet-50	2	0.1	5e-4	0.1	5e-4	0.01	6e-2

Table 2: **Hyperparameters for training the any-precision recognition network**. We use separate sets of learning parameters (learning rate, weight decay) for clipping values of each precision.

Dataset	w_1	w_2	w_3
ActivityNet	0.21	0.5	0.1
FCVID	0.11	1.0	0.1
Mini-Sports1M	0.21	0.5	0.1
Mini-Kinetics	0.21	0.3	0.1

Table 3: **Hyperparameters to train the policy network**.

A. Dataset Details

We evaluate our approach using four standard video recognition benchmark datasets, namely ActivityNet-v1.3 [1], FCVID [4], Mini-Sports1M [5] and Mini-Kinetics [2]. Below we provide more details on each of the dataset.

ActivityNet. We use the v1.3 split of ActivityNet dataset which consists of more than 648 hours of untrimmed videos from a total of 20K videos. Specifically, this dataset has 10,024 videos for training, 4926 videos for validation and 5044 videos for testing with an average duration of 117 seconds. It contains 200 different daily activities such as: walking the dog, long jump, and vacuuming floor. We use the training videos to train our network, and the validation set for testing as labels in the testing set are withheld by the authors. The dataset is publicly available to download at http://activity-net.org/download.html.

FCVID. Fudan-Columbia Video Dataset (FCVID) contains total 91,223 Web videos annotated manually according to 239 categories (45,611 videos for training and 45,612 videos for testing). The categories cover a wide range of topics like social events, procedural events, objects, scenes, etc. that form in a hierarchy of 11 high-level groups (183 classes are related to events and 56 are objects, scenes, etc.). The total duration of FCVID is 4,232 hours with an average video duration of 167 seconds. The dataset is available to download at http://bigvid.fudan.edu.cn/FCVID/.

Mini-Sports1M. Mini-Sports1M is a subset of Sports-1M [5] dataset with 1.1M videos of 487 different fine-grained sports. It is assembled by [3] using videos of length 2-5 mins, and randomly sample 30 videos for each class for training, and 10 videos for each class for testing. The classes are arranged in a manually-curated taxonomy that contains internal nodes such as Aquatic Sports, Team Sports, Winter Sports, Ball Sports, etc, and generally becomes fine-grained by the leaf level. We obtain the training and testing splits from the authors of [3] to perform our experiments. Both training and testing videos in this dataset are untrimmed. This dataset is available to download at https://github.com/gtoderici/sports-1m-dataset.

Mini-Kinetics. Kinetics-400 is a large-scale dataset containing 400 action classes and 240K training videos that are collected from YouTube. Since the full Kinetics dataset is quite large and the original version is no longer available from official site (about $\sim$15% videos are missing), we use the Mini-Kinetics dataset that contains 121K videos for training and 10K videos for testing, with each video lasting 6-10 seconds. We use official training/validation splits of Mini-Kinetics released by authors [6] in our experiments.

Model	mAP (%)	GFLOPs
ActivityNet		
No LSTM	74.1	28.8
LSTM	**74.8**	28.1
Mini-Kinetics		
No LSTM	46.1	26.4
LSTM	**46.4**	26.8

Table 4: **Effect of LSTM on ActivityNet and Mini-Sports1M**.

B. Implementation Details

In this section, we provide more details regarding the implementation. We train the any-precision recognition network from the full-precision recognition network pretrained on the same dataset for 100 epochs. Then we optimize the policy network accompanied with the well-trained (frozen) any-precision recognition network for 50 epochs and the policy network is initialized with the weight pretrained on the same dataset as well. For our experiments, we use 12 NVIDIA Tesla V100 GPUs for training the any-precision recognition network and 6 GPUSs for training the policy network. All our models were implemented and trained via PyTorch. In Table 2 and 3, we provide the initial value (α_{init}), learning rate (α_{lr}) and weight decay (α_{wd}) for each precision to train the any-precision recognition network, as well as hyperparameters w_1, w_2 and w_3 (in Eq. (13) in the main paper) to train the policy network. The data augmentations in our approach are based on the practices in [7]. We first randomly resize the shorter side of an image to a range of [256, 320) while keeping aspect ratio and then randomly crop a 224×224 region and normalize it with the ImageNet's mean and standard deviation to form the input ($16 \times 224 \times 224$). The training time depends on the size of datasets and the task. We will make our code publicly available after the acceptance.

C. Additional Ablation Studies

Effectiveness of LSTM. We investigate the effectiveness of LSTM for modeling video causality in the policy network by comparing with a variant of **VideoIQ** without LSTM (see Table 4). On ActivityNet and Mini-Sports1M datasets, the variant without LSTM yields 0.7% and 0.3% lower mAP with similar GFLOPs than **VideoIQ** respectively. This demonstrates that LSTM is critical for good performance as it makes the policy network aware of all useful information seen so far by aggregating the sequence history.

Effect of Different Losses. Similar to Table 5 of the main paper, we further ablate different losses on Mini-Sports1M (see Table 5) and observe that without knowledge transfer from a pretrained full-precision model, our method only achieves 44.6% with similar amount of GFLOPs. It once again demonstrates the importance of using the full-precision model as the teacher for effective training of lower precisions. When training without efficiency loss (by setting $\mathcal{L}_e = 0$), it achieves 46.6% mAP (0.2% improvement) but with 118% more FLOPs. Furthermore, $\mathcal{L}_b$ and $\mathcal{L}_d$ both improve the performance with similar computational cost.

L_{ce}	L_{kd}	L_e	L_b	L_d	mAP (%)	GFLOPs
✓		✓	✓	✓	44.6	26.5
✓	✓				46.6	58.5
✓	✓	✓			46.3	28.5
✓	✓	✓	✓		46.2	26.9
✓	✓	✓	✓	✓	**46.4**	26.8

Table 5: **Effect of different losses on Mini-Sports1M**.

Decision Space Ω	mAP (%)	GFLOPs
{32, 0}	43.9	28.7
{32, 4, 2}	46.1	29.3
{32, 4, 0}	43.9	33.5
{32, 2, 0}	46.0	32.9
{32, 4, 2, 0}	**46.4**	**26.8**

Table 6: **Effect of different decision space on Mini-Sports1M**.

Effect of Decision Space. Similar to Table 6 in main paper, we show the effect of decision space Ω on Mini-Sports1M (see Table 6). We adjust the training loss to keep their GFLOPS at the same level and we only compare the differences in recognition performances. Only skipping frames yields 43.9% in mAP (0.5% lower than $\Omega = \{32, 4, 2, 0\}$). Among all the alternatives, the best strategy is to set $\Omega = \{32, 4, 2, 0\}$ for achieving top performance of 46.4% in mAP with 26.8 GFLOPs.

D. Qualitative Results

In this section, we provide additional qualitative examples to visualize the learnt policy (see Figure 1). Videos are uniformly sampled in 8 frames. **VideoIQ** processes most informative frames with 32-bit precision while it skips or uses lower precision for the less informative frames without sacrificing accuracy (see top 4 examples in Figure 1: "Swimming", "Tractor Pulling", "Bujinkan" and "Using Segway"). Moreover, it uses 2-bit precision instead of 32-bit precision (see bottom 2 examples in Figure 1: "Riding Camel" and "Freestyle Football") after being confident about the action.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of*

Figure 1: **Qualitative examples.** Our proposed approach **VideoIQ** processes more informative frames with high precision and less informative ones with lower precision or skip them when irrelevant, for efficient video recognition. Best viewed in color.

the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017. 1

[3] Gao, Ruohan and Oh, Tae-Hyun, and Grauman, Kristen and Torresani, Lorenzo. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[4] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 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, 2017. 1

[5] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1

[6] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *ECCV*, 2020. 1

[7] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2