Supplementary Material for DualFormer

Yuxuan Liang[1,2], Pan Zhou[1], Roger Zimmermann[2], and Shuicheng Yan[1]

[1] Sea AI Lab [2] National University of Singapore
{yuxliang,rogerz}@comp.nus.edu.sg {zhoupan,ysc}@sea.com

To support the reproducibility of the results in this study, we have included our code in the supplementary files. In the following sections, we present additional implementation details, more experimental results, and discuss the potential limitations of our proposed approach.

A Additional Implementation Details

In the main body of our paper, we have illustrated the implementation details on the five video datasets. Here, we will introduce the pretraining settings in Sec. A.1 and A.2 and more implementation details of our model in Sec. A.3.

A.1 Initialization by Pretrained Models

Vision Transformer (ViT) [8,15] and its variants [13,6,5,16] have achieved promising results on image classification when trained on large-scale datasets, since transformers lack some of the inductive biases of CNNs. Nevertheless, even the largest video dataset such as Kinetics [10] have much fewer labelled instances than these image counterparts, e.g., ImageNet [7,12], making it extremely challenging to train a video transformer from scratch [3,1].

To tackle this challenge, we follow existing video transformers [3,1,14,2] to utilize a pretrained 2D model (e.g., pretrained on ImageNet) as the initialization of our 3D model. Then, we can fine-tune it on the video datasets for the recognition tasks. In contrast to the pretrained 2D DualFormer, only three building blocks in 3D DualFormer are in different shapes – 1) the patch embedding layer at the first stage; 2) the depth-wise convolution for generating position embeddings in PEG at each stage; and 3) the depth-wise convolution for constructing global priors in GP-MSA. Note that all these three blocks are composed of convolution operations, we thereby employ the common approach [4] for initializing 3D convolution kernels from 2D filters, i.e., we *inflate* the 2D convolution kernels by replicating the filters along the temporal dimension and averaging them.

A.2 Pretraining Settings on ImageNet

When pretraining our DualFormer on ImageNet-1K, we mostly follow the settings of DeiT [15] and Swin [13]. To be more specific, we employ an AdamW optimizer [11] for 300 epochs together with a cosine decay learning rate scheduler and 20 epochs of linear warm-up. The batch size is set to 1024, and the initial learning rate is 0.001. To avoid overfitting, a weight decay rate of 0.05 is used in our method. We include most of the augmentation and regularization strategies of [15] in training, except for repeated augmentation and exponential moving average (EMA), which has been verified ineffective in Swin [13].

For DualFormer-B, we also pretrain it on the larger version of ImageNet, i.e., the ImageNet-21K dataset which contains 14.2 million images and 22 thousand classes. Following Swin Transformer [13], we utilize an AdamW optimizer for 100 epochs using a linear decay learning rate scheduler with a 5-epoch linear warm-up strategy. A batch size of 1024, an initial learning rate of 5e-4, and a weight decay of 0.01 are used. We also employ a stochastic depth drop rate 0.2 to improve its generalization ability.

A.3 Additional Training Details

We implement our models via PyTorch 1.9.0 and mmaction2 which is an open-source toolbox for video understanding. All models are trained on 8 Nvidia A100 GPUs. For example, training a DualFormer-T on Kinetics-400 takes ∼31 hours on 8 A100 GPUs, while training a larger model DualFormer-B on Kinetics-400 requires ∼3 days on 8 A100 GPUs. For the small models (i.e., DualFormer-T/S) on Kinetics-400, we utilize a variant[1] of token labeling [9] as an additional augmentation method to improve their performance, using our DualFormer-B as the teacher model (see Sec B.2 for its effects).

B More Experimental Results

B.1 Results on Image Classification

As our model is pretrained on ImageNet, here we present the experimental results on ImageNet classification in Table 1. We select the most representative pretrained models (DeiT [15] and Swin [13]) of existing video transformers for a comparison. As shown in Table 1, our DualFormer clearly outperforms both baselines with similar computation costs and parameter sizes.

B.2 Effects of Token Labeling

In image processing, token labeling [9] assigns each patch token with an individual location-specific supervision generated by a machine annotator and reformulates the image classification problem into multiple token-level recognition problems. Here, we employ our based model DualFormer-B as the annotator and explore the effects of token labeling on video recognition. As depicted in Table 2, DualFormer-T and DualFormer-S can achieve 0.5% and 0.3% higher top-1 score by using the token labeling trick, respectively. Although these improvements are not very significant compared to the image counterparts [9,17], we believe designing a more powerful video-based token labeling would bring more considerable gains. We leave this for our future work.

[1] MixToken [9] is turned off since it does not work in our experiments.

Method	FLOPs (G)	Param (M)	Top-1 (%)
DeiT-T [15]	4.6	22	79.8
DeiT-B [15]	17.5	86	81.8
Swin-T [13]	4.5	28	81.2
Swin-S [13]	8.7	50	83.1
Swin-B [13]	15.4	88	83.4
DualFormer-T	4.2	22	81.9
DualFormer-S	8.1	50	83.5
DualFormer-B	14.7	88	84.0

Table 1. Results on ImageNet-1K classification. All models are trained and evaluated on 224×224 resolution.

Method	TokenLabel	Top-1 (%)	Top-5 (%)
DualFormer-T		79.0	93.7
DualFormer-T	✓	79.5	94.1
DualFormer-S		80.3	94.5
DualFormer-S	✓	80.6	94.9

Table 2. Effects of token labeling [9] on Kinetic-400.

C Limitations

Even though we have verified the effectiveness of our proposed DualFormer through extensive experiments, our method still has space for improvement in terms of the following two aspects.

Limitation 1 (Pretrained model). Similar to the existing transformer-based architectures [3,2,1,14], our DualFormer is heavily dependent on the pretrained dataset, i.e., ImageNet-1K/21K, resulting in a large additional training time. When training from scratch, all versions of our models suffer 20%∼30% decrease in the top-1 accuracy on Kinetics-400. Hence, how to remove the strong dependency on pretrained models still remains an open problem.

Limitation 2 (FLOPs bottleneck). To improve efficiency, we have investigated the FLOPs of each layer in our DualFormer. We surprisingly find that the major bottleneck is the MLPs for non-linear transformation instead of the MSA computations, since we have greatly reduced the number of keys/values in MSA via our method. For example, DualFormer-T totally requires 60 GFLOPs during inference while ∼50% of the FLOPs are from MLP computations. According to this observation, we have tried to reduce the MLP expansion rate from 4 to $\{1, 2, 3\}$ and train the new models. However, the performance of DualFormer-T degrades by 5%∼12% on the pretrained dataset. This fact verifies the importance of MLPs for non-linear transformation in a large feature space. Then, we also try to factorize the MLP weights into two lower-rank matrix for FLOPs reduction. Although such matrix factorization reduces 25% of the FLOPs, it causes ∼1% drop in the top-1 score on Kinetics-400. Thus, it is still challenging to reduce the FLOPs of MLPs. We will continue to explore this direction.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691 (2021) 1, 3
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021) 1, 3
3. Bulat, A., Perez-Rua, J.M., Sudhakaran, S., Martinez, B., Tzimiropoulos, G.: Space-time mixing attention for video transformer. arXiv preprint arXiv:2106.05968 (2021) 1, 3
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 1
5. Chen, C.F., Panda, R., Fan, Q.: Regionvit: Regional-to-local attention for vision transformers. arXiv preprint arXiv:2106.02689 (2021) 1
6. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. In: NeurIPS 2021 (2021) 1
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 1
9. Jiang, Z., Hou, Q., Yuan, L., Daquan, Z., Shi, Y., Jin, X., Wang, A., Feng, J.: All tokens matter: Token labeling for training better vision transformers. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021) 2, 3
10. 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
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 1
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012) 1
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV) (2021) 1, 2, 3
14. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021) 1, 3
15. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021) 1, 2, 3
16. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122 (2021) 1
17. Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S.: Volo: Vision outlooker for visual recognition. arXiv preprint arXiv:2106.13112 (2021) 2