Supplementary for Recurring the Transformer for Video Action Recognition

Jiewen Yang[1] Xingbo Dong[1,2*] Liujun Liu[1*] Chao Zhang[1]
Jiajun Shen[1] Dahai Yu[1]
[1]TCL Corporate Research (HK) Co., Ltd, [2]Yonsei University, Seoul, South Korea
{jiewen.yang,liujun.liu,chao46.zhang,sjj,dahai.yu}@tcl.com, xingbo.dong@yonsei.ac.kr

1. Conventional ViT and Recurrent ViT for Video Recognition

This section demonstrates the difference between RViT and conventional ViT-based methods (CViT), as depicted in the Figure S1a and Figure S1b. As a batch of frames is feed into the CViT for inference, each frame among the batch is treated equally in both spatial and temporal domain. Processing in batch can preserve long-distance information well among the batch, but also lead to enormous memory consumption. In contrast, our RViT accepts a single frame in each moment while using the hidden state to transfer the information from previous frames, hence significantly reduce the memory consumption.

2. Boundary between Adjacent Actions in a Video

As shown in Figure S2, **there is usually a clear stage boundary between adjacent actions in a video**, e.g., the boundary between sitting and drinking from a cup. There is less dependence between adjacent actions compared with text modality, hence the forget of sitting action won't impair the recognition of drinking action. This also justifies the usage of aggregated temporal features over global-attention-based temporal features.

3. Training Detail

In this section, the training details for different variant of models are discussed, we first introduce the training acceleration, and then discuss the hyper-parameters settings. Source code of this work will be released after publication.

3.1. Training Acceleration

Half-Precision and Mix-Precision Training The half-precision and mix-precision training strategy can efficiently increase the batch size and the training speed. However, the half-precision training might incur overflow or underflow problem. Thus each layers in RViT may need to be adjusted independently to avoid such issues. Here, we use the apex introduced by Nvidia as the prescription, which provide the solution for both half-precision and mix-precision.

Model	Optimizer	Learning Rate	Training Batch	length	Pre-Train
RViT-S°	AdamW	3e-4	32	32	N/A
RViT°	AdamW	3e-4	32	32	N/A
RViT-L°	AdamW	3e-4	32	32	N/A
RViT	SGD	5e-3	32	32	IN-21K
RViT-L	SGD	5e-3	32	32	IN-21K
RViT-XL	SGD	5e-3	16	32	IN-21K

Table S1. **Training Hyper Parameters.** This table shows the configuration of the training hyper parameters for different models.

3.2. Hyper-Parameters

Noted that we use the standard ViT pre-trained model in our experiments to initialize the RViT. Specifically, the spatial domain related weights (W_x^Q, W_x^K, W_x^V) in each block are initialized with the same weights from the corresponding layer in ViT while the temporal domain related weights (W_h^Q, W_h^K, W_h^V) using the orthogonal initialization.

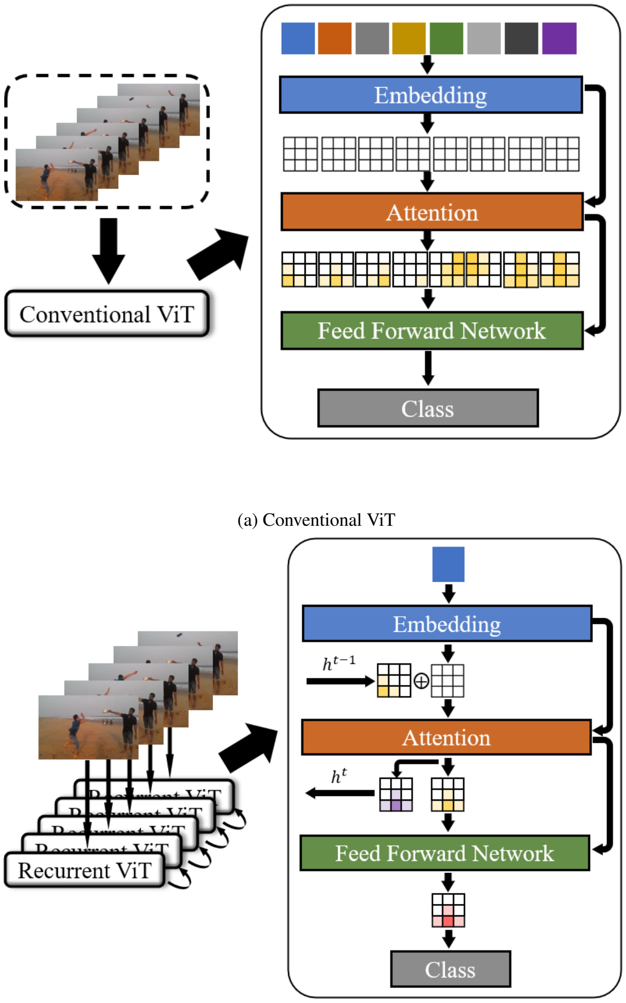
As presented in Table S1, for Kinetics-400 and something-something-V2 datasets, we train our model with public ViT model pre-trained on ImageNet-21K and set the initial learning rate to 5e-3 with synchronized SGD optimizer [1–5]. The learning rate will be divided by 10 for every 10 epochs. For the Jester dataset, we train all the models from scratch. The AdamW optimizer with the learning rate of 3e-4 is adopted in our training. The training batch is all set to 32 except the RViT-XL due to the memory limitation.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?

[*] Work done while interning at TCL Corporate Research (HK) Co., Ltd. and equal contribution



(a) Conventional ViT

(b) Recurrent ViT

Figure S1. **The pipeline of different types of Vision Transformer for video action recognition**. The Figure S1a shows the general recipe of the ViT in video understanding. The Figure S1b illustrates our RViT.

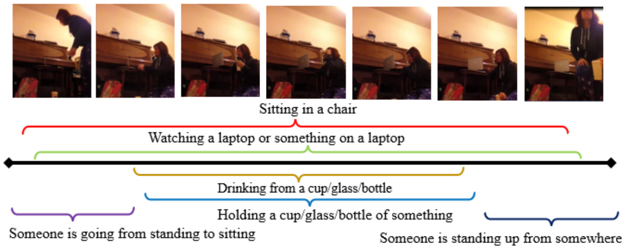

Sitting in a chair
Watching a laptop or something on a laptop
Drinking from a cup/glass/bottle
Holding a cup/glass/bottle of something
Someone is going from standing to sitting
Someone is standing up from somewhere

Figure S2. **Stage boundary among adjacent actions**.

In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 1

[3] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE international conference on computer vision*, 2021. 1

[4] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 1

[5] Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, Jo Henriques, et al. Keeping your eye on the ball: Trajectory attention in video transformers. In *Neural Information Processing Systems*, 2021. 1