Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning Supplemental materials

AJ Piergiovanni Weicheng Kuo Anelia Angelova
Google Research
{ajpiergi,weicheng,anelia}@google.com

1. Implementation Details

Our hyperparameters are summarized in Table 1. For all datasets, we employ random spatial and temporal cropping. For most datasets, these settings were the same. For Charades, we decreased the batch size but used longer, 128 frame clips, as Charades videos are roughly 30 seconds long, compared to 10 seconds for Kinetics.

We also found some training instability when using larger ViT models. When using ViT-L or ViT-H models, we had to decrease the weight decay value as well as the learning rate, otherwise we found the training accuracy dropped to 0 and the loss stayed flat.

For smaller datasets, such as Charades and SSv2, we had to increase the data augmentation settings, as done in previous works, e.g., [9]. We added Mixup and label smoothing and dropout to them.

For all the datasets, we applied RandAugment [1], as we found this to be beneficial. We also kept the number of steps the same for all datasets.

Joint ImageNet and Kinetics Training. When jointly training on the two (or more) datasets, we use a separate fully connected layer to output the class predictions. E.g., for ImageNet and Kinetics-600, we use an FC layer with 1,000 and 600 outputs. We then compute the loss for the relevant head and backpropagate it. During the joint training, we use the same settings as listed in Table 1. We use the joint training for Kinetics 400, 600 and 700. For Charades and SSv2, we use the Kinetics-600+ImageNet pretrained model and finetune it on the dataset.

Full Model Settings. Our model is based on the standard ViT models, thus the core of the approach is the same as previous ViTs [2]. We summarize those settings in Table 2.

In Table 3, we detail the settings for each tube.

2. Additional Experiments on Charades

We include results on Charades [8] to show the effectiveness of this approach on longer videos, since Charades videos are on average 30 seconds long. However, Charades is also a multi-label dataset, and we found it required different settings to effectively train, so we include all those details here.

First, we found that the core multi-tube approach were not performing as well as some prior work (e.g., AssembleNet [6]). Since Charades has a lot of object-related actions and contains longer videos with more temporal information, we modified the core model to make it more suitable for this data. First, we used the interpolation method to increase the tube shapes to:

- $1 \times 16 \times 16$
- $16 \times 16 \times 16$
- $32 \times 8 \times 8$
- $4 \times 32 \times 32$

We note two important factors. First, since we use interpolation to create the larger kernels, the number of learned parameters is the same, and initialized from the same kernels for the other datasets. Second, since the number of strides is unchanged, this results is the same number of tokens. Critically, this change has very little effect on the network and its parameters, but enables the model to better capture the information for Charades.

In Table 4, we report the results. The core MultiTube approach performs quite well, but with the interpolated kernels, is able to perform on par with TokenLearner [5], the state-of-the-art, while still sparsely sampling the video. We also perform similarly using significantly less data, e.g., JFT-300M was used to pre-trained TokenLearner, we accomplish the same performance without such large scale data.

3. Ablations on Tube Shapes.

In Table 5, we provide a detailed study on Kinetics-600 of various tube configurations. We observe that the model isn't overly sensitive to tube shapes, at least on Kinetics-600, but having multiple, different tubes, as well as variation in their shapes is generally beneficial. We use the following tubes in these experiments:

	K400	K600	K700	Charades	SSv2
Optimization					
Optimizer	Adam				
Batch size	256	256	256	64	256
Learning rate schedule	cosine decay + linear warmup				
Linear warmup steps	10,000				
Base learning rate	5e-5 (L,H 1e-5)			1e-3	2e-5
Steps	300,000				
Data augmentation					
Rand augment number of layers [1]	2				
Rand augment magnitude [1]	10				
Weight Decay	0.001 (B) 1e-5 (L,H)				
Mixup [10]	-	-	-	-	0.3
Dropout	-	-	-	0.2	0.3
Label Smoothing	-	-	-	0.1	0.3
Number of Frames	64	64	64	128	32
FPS	15	15	15	6	24

Table 1. Training hyperparamters for our experiments. We note when different settings were used for the base (B), large (L) and huge (H) models.

Model	Layers	Hidden size d	MLP size	Num Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 2. Parameter count for the vit encoder backbones.

(a) $1 \times 16 \times 16$

(b) $4 \times 32 \times 32$

(c) $4 \times 4 \times 4$

(d) $4 \times 12 \times 12$

(e) $8 \times 8 \times 8$

(f) $16 \times 4 \times 4$

(g) $16 \times 16 \times 16$

(h) $32 \times 8 \times 8$

and the following strides:

(i) (4, 16, 16)

(ii) (8, 8, 8)

(iii) (8, 32, 32)

(iv) (16, 16, 16)

(v) (32, 32, 32)

References

[1] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 1, 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 3

[4] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *CVPR*, 2021. 3

[5] Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. 2021. 1, 3

[6] Michael S Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. In *European Con-*

Kernel	Stride	Offset	S2D	params
$8 \times 8 \times 8$	$(16, 32, 32)$	(0, 0, 0)	2x temporal	$512d$
$16 \times 4 \times 4$	$(6, 32, 32)$	(4, 8, 8)	4x spatial	$256d$
$4 \times 12 \times 12$	$(16, 32, 32)$	(0, 16, 16)	-	$576d$
$1 \times 16 \times 16$	$(32, 16, 16)$	(0, 0, 0)	-	$256d$

Table 3. Configuration for the tubes using the in main Tube-ViT. We also report the number of params used by each tube, which depends on d, the hidden size of the ViT model used. The tubes add an additional 1-3M params, depending on the model, a small fraction of the total model size.

	mAP
SlowFast [3]	45.2
AssembleNet-101 [7]	58.6
AssembleNet++-50 [6]	59.8
MoViNet-A6 [4]	63.2
TokenLearner [5]	66.3
MultiTube Tube-ViT-L	61.8
Interpolated TubeViT-L	66.2

Table 4. Charades classification.

Tube Config	K600
(a+iv) + (b+v) + (f+iv)	87.9
(c+iv) + (e+v) + (g+iv)	87.5
(a+iv) + (e+v) + (g+iv)	87.8
(b+iv) + (e+v) + (g+iv)	87.7
(a+iv) + (d+v) + (e+iv) + (h+v)	88.6
(a+iv) + (b+v) + (c+iv) + (h+v)	87.9
(a+iv) + (d+v) + (e+iv) + (f+v)	88.9

Table 5. Ablation on different tube shapes, trained for 50k steps.

ference on Computer Vision, pages 654–671. Springer, 2020. 1, 3

[7] Michael S Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. In *ICLR*, 2019. 3

[8] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 1

[9] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 1

[10] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2