ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning

香港中文大學 The Chinese University of Hong Kong
UNIVERSITY OF SURREY

Junting Pan*, Ziyi Lin*, Xiatian Zhu, Jing Shao, Hongsheng Li

NEURAL INFORMATION PROCESSING SYSTEMS

Models & source code:
https://github.com/linziyi96/st-adapter

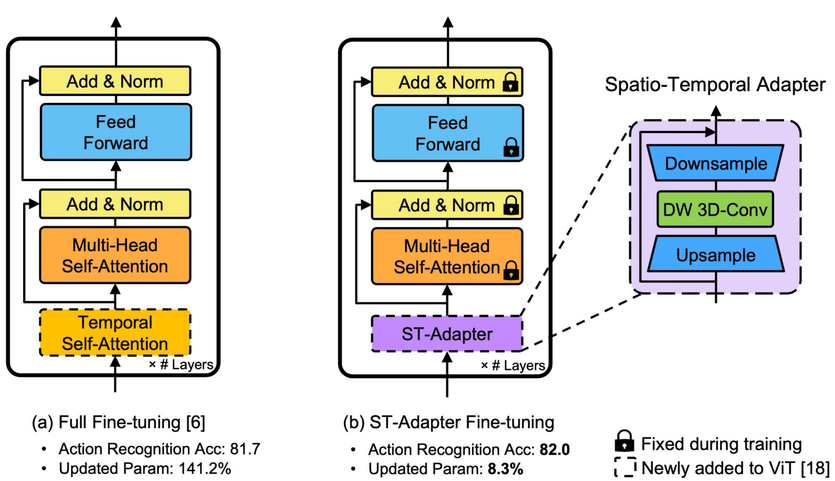
1. Efficient Image-to-Video Transfer Learning

Video pre-training is more challenging
- Data: difficult to collect, store and manage
- Training: long and compute more hungry

Our goal:
- Efficiently adapting pre-trained image models for video tasks.

2. ST-Adapter

$$\texttt{ST-Adapter}(\mathbf{X}) = \mathbf{X} + f\Big(\texttt{DWConv3D}(\mathbf{X}\mathbf{W}_{down})\Big)\mathbf{W}_{up},$$



(a) Full Fine-tuning [6]
- Action Recognition Acc: 81.7
- Updated Param: 141.2%

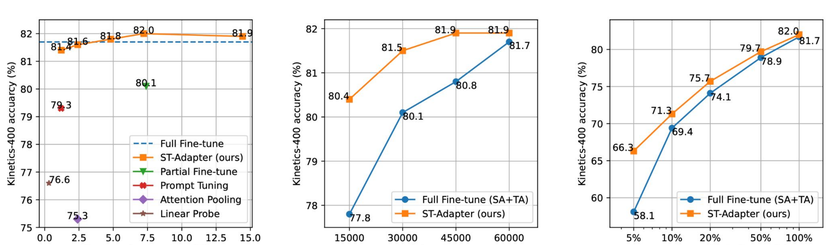
(b) ST-Adapter Fine-tuning
- Action Recognition Acc: **82.0**
- Updated Param: **8.3%**

3. Cross-Modality Fine-Tuning Benchmark

Fine-tuning Methods	Architecture	TM?	Fine-tuned Params (M)	CLIP K400	CLIP SSv2	ImageNet-21K K400	ImageNet-21K SSv2
Full Fine-tuning	SA		86.11	81.0	44.0	76.9	40.0
	SA + TA [6]	✓	121.57	81.7	66.1	78.0	59.5
	SA + TS [9]	✓	93.79	78.0	62.0	**78.5**	**64.4**
Partial Fine-tuning	SA		7.40	80.1	37.6	61.7	20.4
	SA + TA	✓	10.36	80.3	57.5	63.1	29.3
Temporal Fine-tuning	SA + TA	✓	35.8	81.3	59.4	76.5	51.9
Prompt Tuning	SA		1.18	79.3	39.3	71.4	26.3
Attentional Pooling	SA	✓	2.36	75.3	21.5	59.1	15.1
Linear Probe	SA		0.31	76.6	21.9	60.1	14.8
Adapter [27]	SA		6.77	81.6	46.2	76.2	40.5
ST-Adapter (ours)	SA	✓	7.20	**82.0**	**66.3**	76.6	62.8

The *TM?* column shows whether the method includes temporal modelling.

4. Ablation Study on Efficiency



The same ViT-B/16 with CLIP pre-training is used for all experiments.

5. Results

Comparison on Kinetics 400

Model	Pretrain	#Frames	GFlops	Top-1	Top-5
Methods with full-finetuning					
TimeSformer-L[6]	IN21K	96×3×1	7140	80.7	94.7
X-ViT[9]	IN21K	16×3×1	850	80.2	94.7
Mformer-HR[48]	IN-21K	16×3×10	28764	81.1	95.2
MViT-B,32×3[20]	-	32×1×5	850	80.2	94.4
ViViT-L[2]	JFT300M	16×3×4	17352	82.8	95.3
Swin-L(384)[44]	IN21K	32×5×10	105350	84.9	96.7
UniFormer-B[38]	IN1K	32×1×4	1036	82.9	95.4
VATT-Large(320)[1]	HowTo100M	32×3×4	29800	82.1	95.5
TokenLearner[56]	JFT300M	64×3×4	48912	85.4	96.3
ViT-B w/o ST-Adapter	CLIP	8×3×1	419	81.0	95.5
ViT-L w/o ST-Adapter	CLIP	8×3×1	1941	85.8	97.2
Methods with frozen backbone					
Our ViT-B w/ ST-Adapter	CLIP	8×3×1	455	82.0	95.7
Our ViT-L w/ ST-Adapter	CLIP	8×3×1	2062	**86.7**	**97.5**

Comparison on Something-Something v2

Model	Pretrain	#Frames	GFlops	Top-1	Top-5
Methods with full-finetuning					
TimeSformer-HR[6]	IN21K	16×3×1	5109	62.5	-
X-ViT[9]	IN21K	32×3×1	1270	65.4	90.7
Mformer-L[48]	IN21K+K400	32×3×1	3555	68.1	91.2
ViViT-L[2]	IN21K+K400	16×3×4	11892	65.4	89.8
MViT-B-24,32×3[20]	K600	32×1×3	708	68.7	91.5
Swin-B[44]	IN21K+K400	32×3×1	963	69.6	92.7
UniFormer-B[38]	IN1K+K600	32×3×1	777	71.2	92.8
ViT-B w/o ST-Adapter	CLIP	8×3×1	419	44.0	77.0
ViT-L w/o ST-Adapter	CLIP	8×3×1	1941	48.7	77.5
Methods with frozen backbone					
Our ViT-B w/ ST-Adapter	CLIP	8×3×1	489	67.1	91.2
Our ViT-L w/ ST-Adapter	CLIP	8×3×1	2062	**70.0**	**92.3**

References:

[6] Bertasius et.al. Is Space-Time Attention All You Need for Video Understanding? ICML 2021

[18] Dosovitskiy et. al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021