Paper ID 568

Token Shift Transformer for Video Classification

Hao Zhang, Yanbin Hao*, Chong-Wah Ngo

ACM multimedia Chengdu, China OCT 20-24 2021

• Motivation

➢ The attention in Transformers suffers from a heavier computation burden in processing 3D video than in 1/2D language and image.

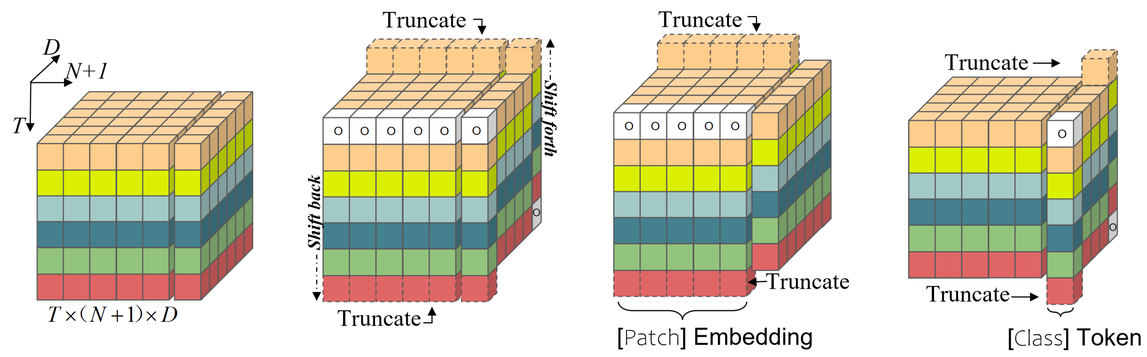
➢ The amount of pair-wise distance calculation of an attention for the Video and Image tensor.

	Tensor Shape	Distance Calculations
Image	$z_i \in \mathbb{R}^{(N+1)\times D}$	$\frac{(N+1)^2}{2}$
Video	$z_v \in \mathbb{R}^{T\times(N+1)\times D}$	$T^2 \cdot \frac{(N+1)^2}{2}$

➢ Spatio-temporal video sequences will introduce an exponential ***explosion of computations***.

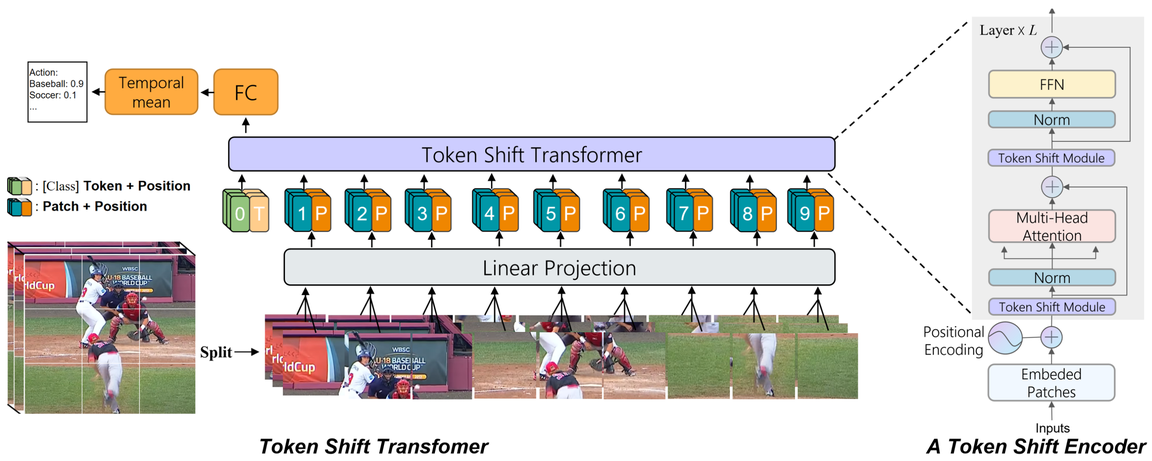
• Proposed Framework

➢ ***Token Shift Module***: a novel zero-parameter/FLOPs operator for modeling temporal relations within each Transformer encoder. Below shows the *TokShift* and other *Shift* variants:



(a) Video Tensor *(b) Temporal Shift* *(c) Patch Shift* *(d) Token Shift*

➢ ***Token Shift Transfomer***: densely plug the TokShift module into a vanilla 2D ViT [1] to build the TokShift-xfmr).



Token Shift Transfomer ***A Token Shift Encoder***

➢ ***Pair-wise distance calculation*** of attention in TokShift-xfmr is reduced from:

$$T^2 \cdot \frac{(N+1)^2}{2} \rightarrow T \cdot \frac{(N+1)^2}{2}$$

• Experiments

➢ ***Ablation Study***:

1. Non-Shift *vs* TokShift

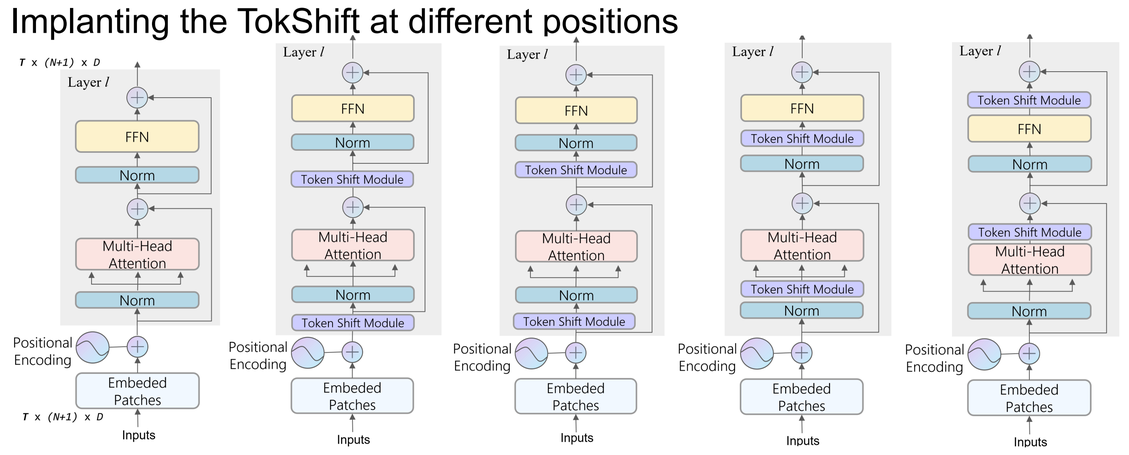
Acc1 (Xfmr \ $T \times S$)	8×8	8×16	8×32	8×64
ViT (video)[9]	76.17	76.32	76.02	75.73
TokShift	76.90 (0.73↑)	76.81 (0.49↑)	**77.28** (1.26↑)	76.60 (0.87↑)

Table 1: NonShift *vs* TokShift under diffrent sampling strategies ("T/S" refers to frames/sampling-step; Accuracy-1).

2. TokShift *vs* Shift variants

Shift Type	Words Shiftted	$GFLOPs \times Views$	Acc1 (%)	Acc5 (%)
ViT (video) [9]	None	134.7×30	76.02	92.52
TemporalShift	Token + Patches	134.7×30	72.88	91.24
PachShift	Patches	134.7×30	73.08	91.17
TokShift	Token	134.7×30	**77.28**	**92.91**

3. Implanting the TokShift at different positions



(a) ViT-Video *(b) TokShift* *(c) TokShift-A* *(d) TokShift-B* *(e) TokShift-C*

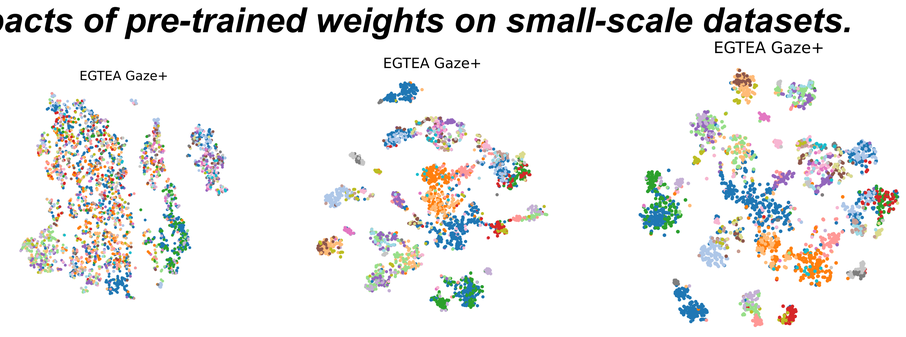
Model	Words Shiftted	$GFLOPs \times Views$	Acc1 (%)	Acc5 (%)
ViT (video) [9]	None	134.7×30	76.02	92.52
TokShift-*A*	Token	134.7×30	76.85	93.10
TokShift-*B*	Token	134.7×30	77.21	92.81
TokShift-*C*	Token	134.7×30	77.00	92.92
TokShift	Token	134.7×30	**77.28**	92.91

4. Comparison to state-of-the-arts on Kinetics-400 Val

Model	Backbone	Pretrain	Inference Res ($H \times W$)	# Frames/Clip T	GFLOPs×Views	Params	Accuracy-1 (%)	Accuracy-5 (%)
I3D [4] from [11]	InceptionV1	ImageNet	224×224	250	108× NA	12M	71.1	90.3
Two-Stream I3D [4] from [11]	InceptionV1	ImageNet	224×224	500	216× NA	25M	75.7	92.0
S3D-G [39]	InceptionV1	ImageNet	224×224	250	71.3×NA	11.5M	74.7	93.4
Two-Stream S3D-G [39]	InceptionV1	ImageNet	224×224	500	142.6×NA	11.5M	77.2	93.0
Non-Local R50 [33] from [11]	ResNet50	ImageNet	256×256	128	282×30	35.3M	76.5	92.6
Non-Local R101[33] from [11]	ResNet101	ImageNet	256×256	128	359×30	54.3M	77.7	93.3
TSM [20]	ResNet50	ImageNet	256×256	8	33×10	24.3M	74.1	91.2
TSM [20]	ResNet50	ImageNet	256×256	16	65×10	24.3M	74.7	-
TSM [20]	ResNext101	ImageNet	256×256	8	NA×10	-	76.3	-
SlowFast 4 × 16 [12]	ResNet50	None	256×256	32	36.1×30	34.4M	75.6	92.1
SlowFast 8 × 8 [12]	ResNet50	None	256×256	32	65.7×30	-	77.0	92.6
SlowFast 8 × 8 [12]	ResNet101	None	256×256	32	106×30	53.7M	77.9	93.2
SlowFast 8 × 8 [12]	ResNet101+NL	None	256×256	32	116×30	59.9M	78.7	93.5
SlowFast 16 × 8 [12]	ResNet101+NL	None	256×256	32	234×30	59.9M	79.8	93.9
X3D-M [11]	X2D [11]	None	256×256	16	6.2×30	3.8M	76.0	92.3
X3D-L [11]	X2D [11]	None	356×356	16	24.8×30	6.1M	77.5	92.9
X3D-XL[11]	X2D [11]	None	356×356	16	48.4×30	11M	79.1	93.9
X3D-XXL [11]	X2D [11]	None	NA	NA	194.1×30	20.3M	**80.4**	94.6
ViT (Video)[9]	Base-16	ImageNet-21k	224×224	8	134.7×30	85.9M	76.02	92.52
TokShift	Base-16	ImageNet-21k	224×224	8	134.7×30	85.9M	77.28	92.91
TokShift (MR)	Base-16	ImageNet-21k	256×256	8	175.8×30	85.9M	77.68	93.55
TokShift (HR)	Base-16	ImageNet-21k	384×384	8	394.7×30	85.9M	78.14	93.91
TokShift	Base-16	ImageNet-21k	224×224	16	269.5×30	85.9M	78.18	93.78
TokShift-Large (HR)	Large-16	ImageNet-21k	384×384	8	1397.6×30	303.4M	79.83	94.39
TokShift-Large (HR)	Large-16	ImageNet-21k	384×384	12	2096.4×30	303.4M	**80.40**	94.45

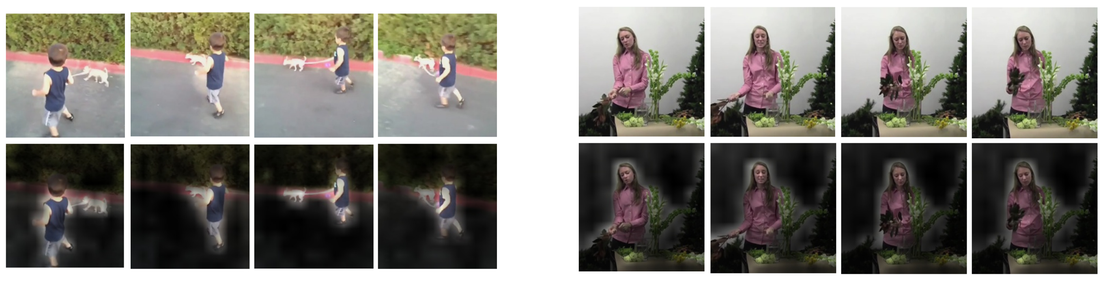
• Visualization

➢ ***Impacts of pre-trained weights on small-scale datasets.***

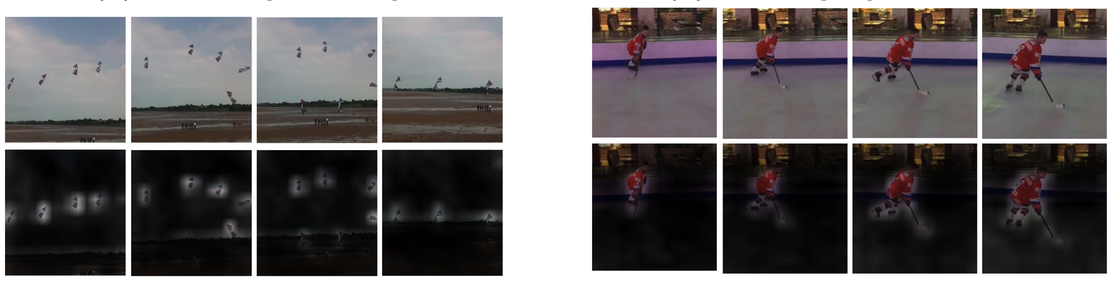


(a) None → EG *(b) ImageNet → EG* *(c) Kinetics400 → EG*

➢ ***Attention Visualization***



(a) "Walking the dog" *(b) "Arranging flowers"*



(c) "Flying kite" *(d) "Playing ice hockey"*

• Conclusion and Resources

➢ ***Conclusions***:

1. Transformer shows architecture universalism for language/image/video.
2. Zero Parameter/FLOPs Shift is generalizable to Transformers.
3. TokShift-xfmr achieves comparable or better performance than SOTA CNN

➢ ***Contact & Resources***

zhanghaoinf@gmail.com
haoyanbin@hotmail.com
cwngo@smu.edu.sg

GitHub PDF Adobe

[1]. "An image is worth 16x16 words transformers for image recognition at scale", ICLR-2021