Prune Spatio-temporal Tokens by Semantic-aware Temporal Accumulation

Shuangrui Ding, Peisen Zhao, Xiaopeng Zhang, Rui Qian, Hongkai Xiong, Qi Tian

Shanghai Jiao Tong University · HUAWEI

ICCV23 PARIS

Abstract:

- **Task:** Efficient Action Recognition
- **Problem:** Increased Complexity in Video Transformer due to the extra temporal dimension.
- **Our solution:** We propose **S**emantic-aware **T**emporal **A**ccumulation score (STA) to prune spatio-temporal tokens integrally. We consider two factors to form STA score: temporal redundancy and semantic importance.

Method（STA）:

Temporal Redundancy

We introduce the accumulative temporal score **A** to model the probability of dropping a token. Next, we eliminate r tokens with the highest scores from **A** at t-th frame and transfer the remaining probability distribution to the next frame via the transition:

$$\mathbf{A}_{t+1} := \mathbb{P}_{\text{drop}}(\mathbf{X}_{t+1}|\mathbf{X}_t^{'})\mathbf{A}_t^{'},$$
$$\mathbb{P}_{\text{drop}}(\mathbf{X}_{t+1}|\mathbf{X}_t^{'}) := \text{softmax}(f(\mathbf{X}_{t+1})f(\mathbf{X}_t^{'})^T),$$

This formulation allows us to aggregate potential redundancy from the first frame to every next frame.

Semantic Importance.

We devise to treat each token differently via its contribution to the semantics of the class. Specifically, we define the semantic score for token $X_{t,s}$ as:

$$\mathcal{F}(\mathbf{X}_{t,s}) = \sum_{i=1}^{d} |\mathbf{X}_{t,s,i}| \in \mathbb{R}^{+}$$

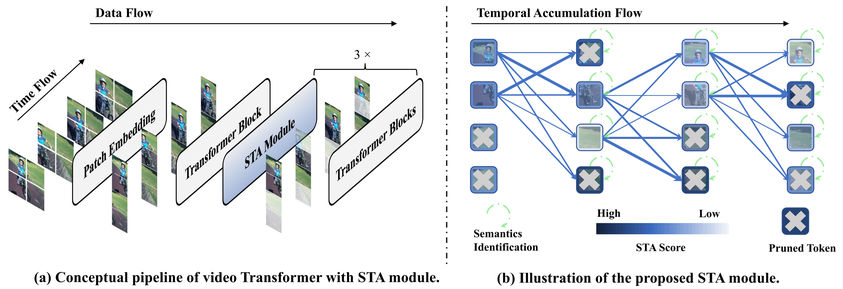
Through the summation of absolute activation values over channel dimension, a high absolute activation tend to represent discriminative category information.

Semantic-aware Temporal Accumulation Score

$$\widetilde{\mathbf{A}}_{t,s} = (1 - \mathcal{F}(\mathbf{X}_{t,s}))\mathbf{A}_{t,s},$$

Three Advantages of STA

- Motion-aware and suitable for video data through temporal accumulation mechanism;
- A simple plug-in module without the introduction of additional parameters and the retraining of the video Transformer;
- Negligible additional FLOPs and feasible for the bulk of computation to be done in parallel.



(a) Conceptual pipeline of video Transformer with STA module. (b) Illustration of the proposed STA module.

Fig 1. An overview of our STA token-pruning algorithm for video Transformers.

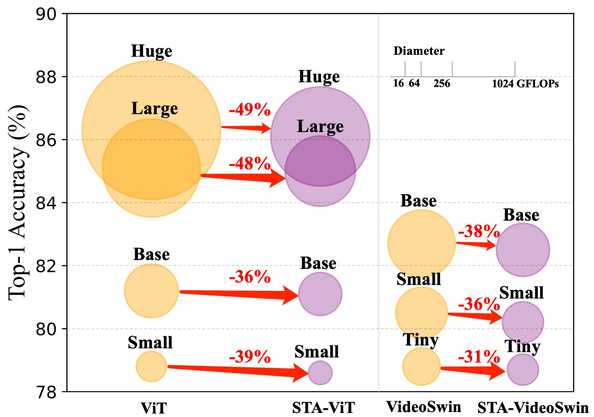


Fig 2. Kinectics-400 result for ViT and VideoSwin. The bubble's area is proportional to FLOPs of a variant in a model family.

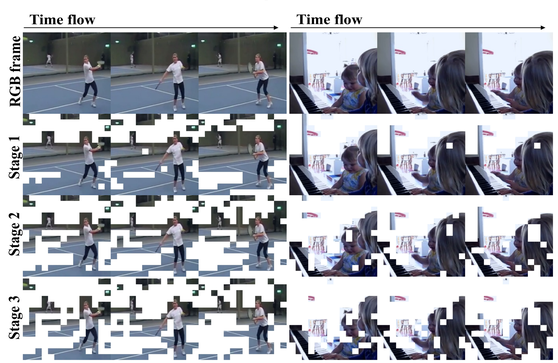


Fig 3. Visualization of the proposed STA strategy. Selected tokens represent diverse and significant parts of a video.

Experiments:

- Figure 2 displays the intuitive bubble plot for our method. The proposed STA saves over 30% FLOPs for all model variants with a negligible drop in performance.
- Table 1 and 2 report our main experimental result on K400 and SSV2, two standard action recognition benchmarks using two mainstream Transformer backbone.
- Tables 3 and 4 offer comparisons with prior research. The proposed STA demonstrates an optimal balance between FLOPs and accuracy.
- We vary the prune number r and compare our STA against the random pruning baseline, as shown in Figure 4.
- Figure 5 reveals sparse patterns, indicating that most tokens do not contribute significantly to the final prediction.

Base Model	Metrics	Drop Number r_1			
		0	32	48	64
ViT-S	K400 Acc. (%)	78.8	78.8 (-0.0)	78.6 (-0.2)	78.1 (-0.7)
	SSV2 Acc. (%)	66.8	66.6 (-0.2)	66.4 (-0.4)	65.8 (-1.0)
	GFLOPs	57	42 (-26%)	35 (-39%)	29 (-49%)
ViT-B	K400 Acc. (%)	81.2	81.2 (-0.0)	81.1 (-0.1)	80.8 (-0.4)
	SSV2 Acc. (%)	70.6	70.4 (-0.2)	70.3 (-0.3)	69.9 (-0.7)
	GFLOPs	180	136 (-24%)	116 (-36%)	96 (-47%)
ViT-L	K400 Acc. (%)	85.1	85.2 (+0.1)	85.1 (-0.0)	85.0 (-0.1)
	GFLOPs	597	446 (-25%)	376 (-37%)	308 (-48%)
ViT-H	K400 Acc. (%)	86.3	86.3 (-0.0)	86.2 (-0.1)	86.1 (-0.2)
	GFLOPs	1192	890 (-25%)	748 (-37%)	611 (-49%)

Tab 1. Main Results for STA-ViT on K400 and SSV2.

Base Model	Metrics	Drop Number r_1			
		0	192	256	320
VideoSwin-T	K400 Acc. (%)	78.8	78.7 (-0.1)	78.7 (-0.1)	78.6 (-0.2)
	GFLOPs	88	68 (-23%)	61 (-31%)	54 (-39%)
VideoSwin-S	K400 Acc. (%)	80.5	80.3 (-0.2)	80.2 (-0.3)	80.1 (-0.4)
	GFLOPs	166	121 (-27%)	106 (-36%)	91 (-45%)
VideoSwin-B	K400 Acc. (%)	82.7	82.5 (-0.2)	82.5 (-0.2)	82.3 (-0.4)
	K400 GFLOPs	282	202 (-28%)	176 (-38%)	149 (-47%)
	SSV2 Acc. (%)	69.6	69.6 (-0.0)	69.5 (-0.1)	69.2 (-0.4)
	SSV2 GFLOPs	321	241 (-25%)	215 (-33%)	188 (-41%)

Tab 2. Main Results for STA-VideoSwin on K400 and SSV2.

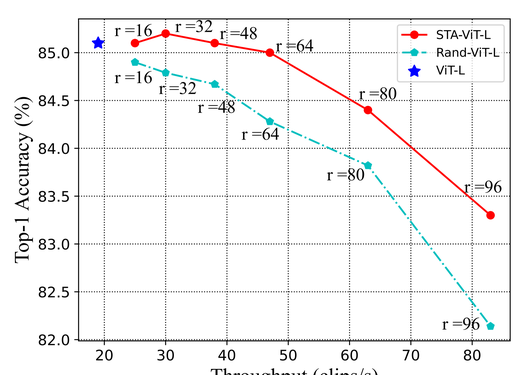


Fig 4. Top-1 accuracy and throughput between ours and random-drop with various prune numbers r.

Model	GFLOPs×views	Top-1
TimeSformer-L [2]	8353 × 1 × 3	80.7
Motionformer-L [28]	1185 × 10 × 3	80.2
ViViT [1]	3981 × 4 × 3	84.9
Swin-L [23]	2107 × 10 × 5	84.9
MViTv2-L [19]	2828 × 5 × 3	86.1
ViT-H [35]	1192 × 5 × 3	86.3
STTS-VideoSwin-B [40]	253 × 4 × 3	81.9
ToMe-ViT-L [3]	281 × 10 × 1	84.5
STA^{320}-VideoSwin-B (ours)	149 × 4 × 3	82.3
STA^{64}-ViT-L (ours)	308 × 5 × 3	85.0
STA^{64}-ViT-H (ours)	611 × 5 × 3	86.1

Tab 3. Comparisons with prior arts on K400.

Model	GFLOPs×views	Top-1
TimeSformer-L [2]	5549 × 1 × 3	62.4
Motionformer-L [28]	1185 × 1 × 3	68.1
MViTv2-B [19]	225 × 1 × 3	70.5
VideoSwin-B [23]	321 × 1 × 3	69.6
ViT-B [35]	180 × 2 × 3	70.6
STTS-VideoSwin-B [40]	237 × 1 × 3	68.7
STA^{320}-VideoSwin-B (ours)	188 × 1 × 3	69.2
STA^{48}-ViT-B (ours)	116 × 2 × 3	70.3

Tab 4. Comparisons with prior arts on SSV2.

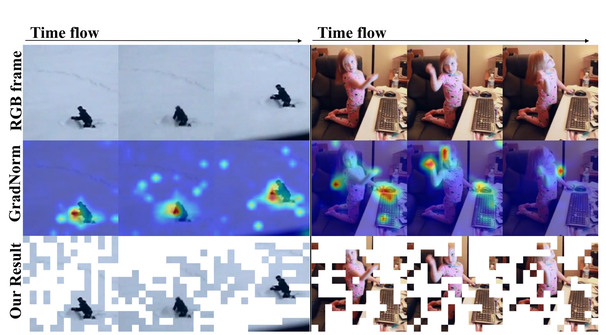


Fig 5. Gradient visualization for ViT-Large on the Kinetics-400 validation set. Our pruning algorithm preserves the area of rich semantics well.