NEURAL INFORMATION PROCESSING SYSTEMS

THE UNIVERSITY of ADELAIDE

OnlineTAS: An Online Baseline for Temporal Action Segmentation

NUS National University of Singapore

Qing Zhong[1,2,*], Guodong Ding[2,*], Angela Yao[2]

[1]University of Adelaide, [2]National University of Singapore,
[*]Equal Contribution.

Temporal Action Segmentation

Offline

- Complete video used in inference

Temporal Action Segmentation

Offline

- Complete video used in inference

Online

- Clip or single frame used in inference.

Adaptive Memory Bank

- Fixed size memory length w
- Long-term size increasing, but not exceeding 2/3 w.
- Short-term size decreasing.

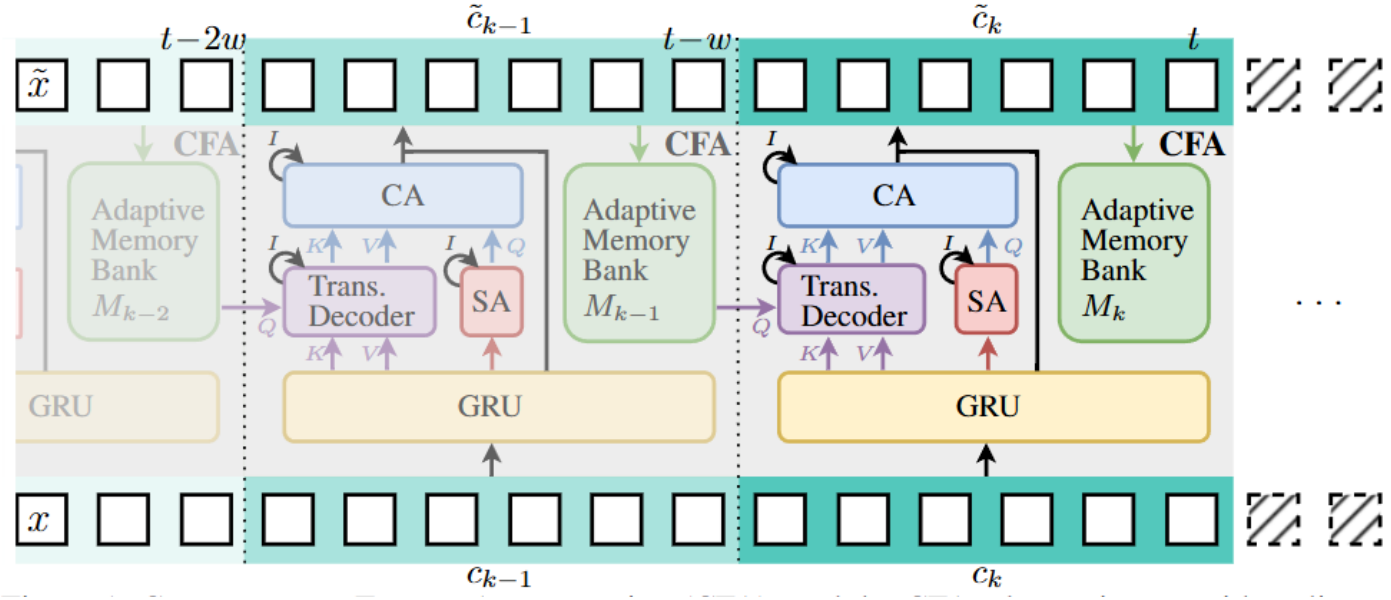
Algorithm 1 Adaptive Memory Update

Require: $\{c_k\}_{k=1}^{K}$, w

Initialize $M_0^{\text{short}} \leftarrow c_1, M_0^{\text{long}} \leftarrow \varnothing$
for $k \in [1...K]$ **do**
    $\tilde{c}_k = \text{CFA}(c_k, M_{k-1})$
    $m_k = \text{Conv1D}(\tilde{c}_k)$
    if $\text{len}(M_{k-1}^{\text{long}}) \leq \frac{2}{3}w$ **then**
        $M_k^{\text{long}} = \text{concat}(M_{k-1}^{\text{long}}, m_k)$
    else
        $M_k^{\text{long}} = \text{concat}(M_{k-1}^{\text{long}}[1:], m_k)$
    end if
    $M_k^{\text{short}} = \tilde{c}_{k-1}[\text{len}(M_k^{\text{long}}):]$
    $M_k = [M_k^{\text{long}}, M_k^{\text{short}}]$
end for

Context-aware Feature Augmentation (CFA)

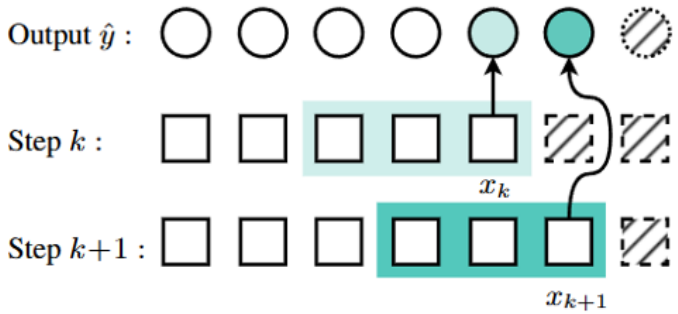
- Capture video information.
- Achieve a more effective memory.
- Information exchange with local clip window
- Combine features.



Inference

Two inference mode

- Online inference

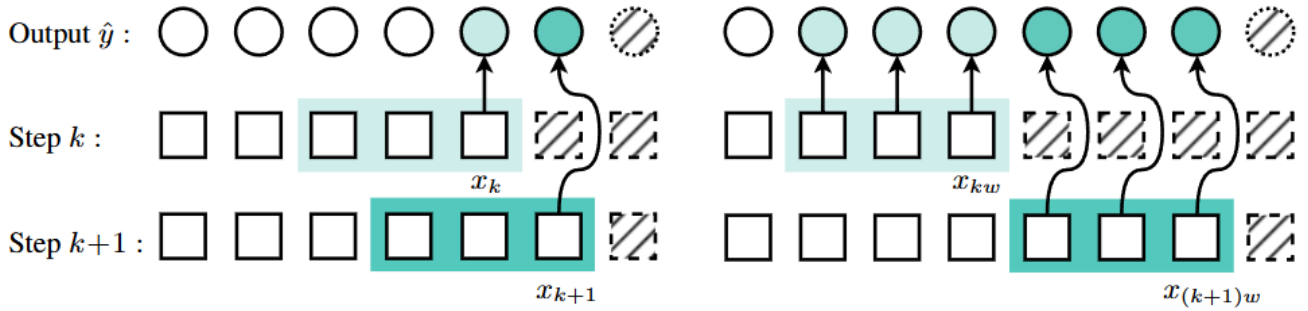


(a) Online Inference ($\delta = 1$)

Inference

Two inference mode

- Online inference
- Semi-online inference



(a) Online Inference ($\delta = 1$) (b) Semi-online Inference ($\delta = w$)

Post-processing

- Selecting valid action segment
- Mitigating the over-segmentation

Algorithm 2 Post-processing for Online TAS

Compute $\ell_{\text{min}} = \sigma \times T_{\text{max}}$
Initialize $\ell = 0$
for each frame t **do**
    if $q_t < \theta$ and $\ell < \ell_{\text{min}}$ **then**
        $\hat{y}_t^* = \hat{y}_{t-1}^*$
        $\ell = \ell + 1$
    else
        $\hat{y}_t^* = \hat{y}_t$
        $\ell = 0$
    end if
end for

Ablation study of module components

- The GRU and Mem. has ability to accumulate context information
- The CFA module enhanced clip-wise features with GRU and Mem. improves performance

GRU	CFA	Mem.	Acc	Edit	F1 @ {10, 25, 50}		
-	-	-	75.2	19.6	26.8	24.4	19.6
✓	-	-	78.1	27.1	37.9	34.7	26.7
-	✓	-	76.2	22.3	30.1	27.0	21.9
✓	✓	-	79.1	29.0	38.5	35.5	28.3
-	✓	✓	78.9	29.2	38.7	35.1	28.8
✓	✓	✓	**82.4**	**32.8**	**43.0**	**41.1**	**34.7**

Ablation study of memory composition

- Each type of memory contributes to performances improvements
- Both long and short memory information are equally important.

M^{short}	M^{long}	Acc	Seg.
✓	-	80.3	36.7
-	✓	80.4	36.4
✓	✓	**82.4**	**37.9**

Comparison with SOTA methods on three benchmarks

	Method	GTEA [12]					50Salads [38]				
		Acc	Edit	F1 @ {10, 25, 50}			Acc	Edit	F1 @ {10, 25, 50}		
offline	MS-TCN [11]	78.7	84.0	88.3	86.6	72.8	81.2	65.8	72.8	70.4	61.7
	MS-TCN + p.p.	78.7	85.2	89.6	88.3	73.3	80.4	74.1	82.0	79.2	70.2
	ASFormer [45]	79.7	84.6	90.1	88.8	79.2	85.6	79.6	85.1	83.4	76.0
	DiffAct [25]	82.2	89.6	92.5	91.5	84.7	87.4	88.9	90.1	89.2	83.7
online	LSTR [44]	63.7	33.2	41.5	37.7	25.0	60.5	5.0	8.2	6.6	4.1
	Causal TCN	74.4	66.6	73.9	70.3	57.2	75.2	19.6	26.8	24.4	19.6
	Ours^{online}	75.8	66.8	74.3	71.5	60.3	79.1	29.0	38.5	35.5	28.3
	Ours^{online} + p.p.	73.5	75.4	80.3	76.9	66.6	76.7	69.2	73.1	70.5	62.8
	Ours^{semi}	**77.1**	68.1	76.7	73.5	63.9	**82.4**	32.8	43.0	41.1	34.7
	Ours^{semi} + p.p.	76.0	**79.7**	**84.9**	**81.4**	**69.2**	79.4	**75.0**	**82.5**	**80.2**	**68.0**

Table 8: Comparison with the state-of-the-art methods on GTEA and 50Salads.

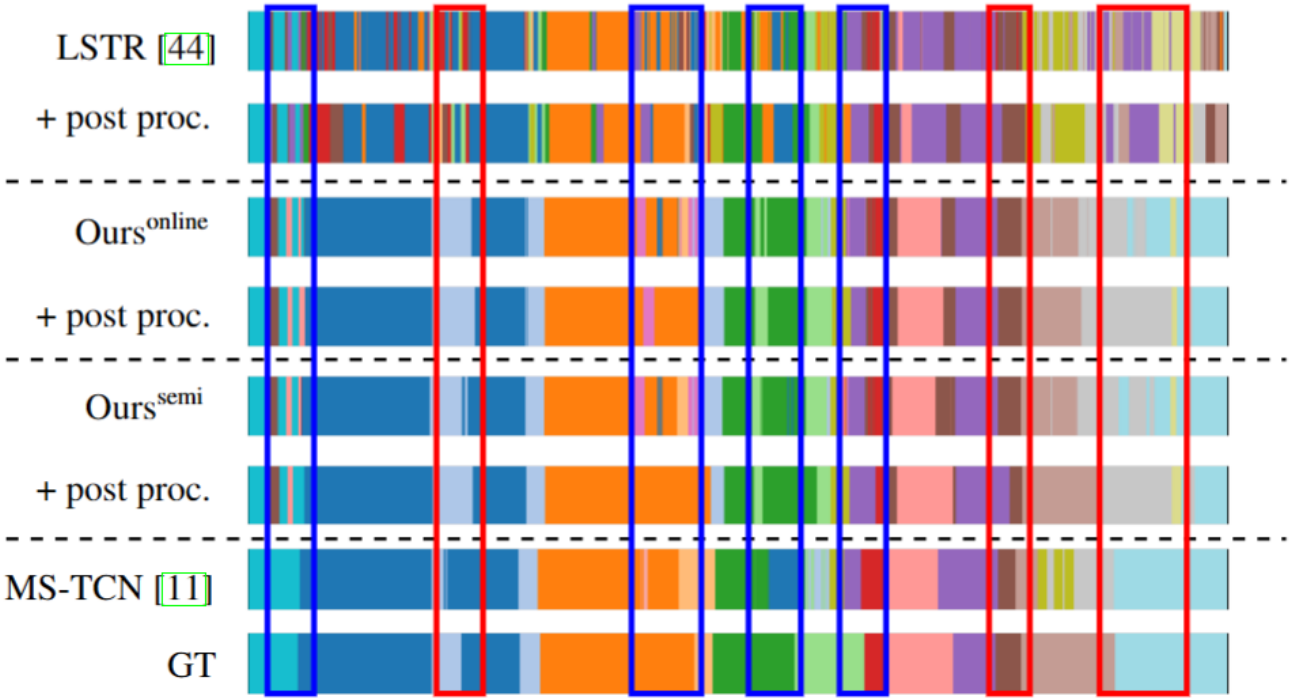
	Method	Breakfast [18]				
		Acc	Edit	F1 @ {10, 25, 50}		
offline	MS-TCN [11]	69.3	67.3	64.7	59.6	47.5
	ASFormer [45]	73.5	75.0	76.0	70.6	57.4
	DiffAct [25]	75.1	76.4	80.3	75.9	75.1
online	MV-TAS [13]	41.6	-	-	-	-
	LSTR [44]	24.2	4.9	5.5	3.9	1.7
	Causal TCN	55.3	18.7	15.1	11.7	8.3
	Ours^{online}	56.7	19.3	16.8	13.9	9.3
	Ours^{online} + p.p.	52.9	55.7	54.8	45.8	30.5
	Ours^{semi}	**57.4**	19.6	17.8	14.8	10.1
	Ours^{semi} + p.p.	53.8	**57.5**	**56.4**	**47.3**	**31.4**

Table 9: Comparison with the state-of-the-art methods on Breakfast.

Our method: SOTA in online setting, and comparable with offline setting.

Visualization

- Semi-online inference producing smoother predictions.
- Removes short fragments (blue boxes).
- Reduce accuracy near action boundaries (red boxes).



Take aways

Temporal Interaction

- Interaction between the current video clip features and the past memory bank is essential for achieving good performance.

Inference mode

- Semi-online inference, which retains dense predictions generated from all frames as the final output, yields better performance.

Post-processing Assist

- Post-processing is effective in addressing over-segmentation.

Thank you