Real-time Online Video Detection with Temporal Smoothing Transformers

Yue Zhao, Philipp Krähenbühl

{yzhao, philkr}@cs.utexas.edu

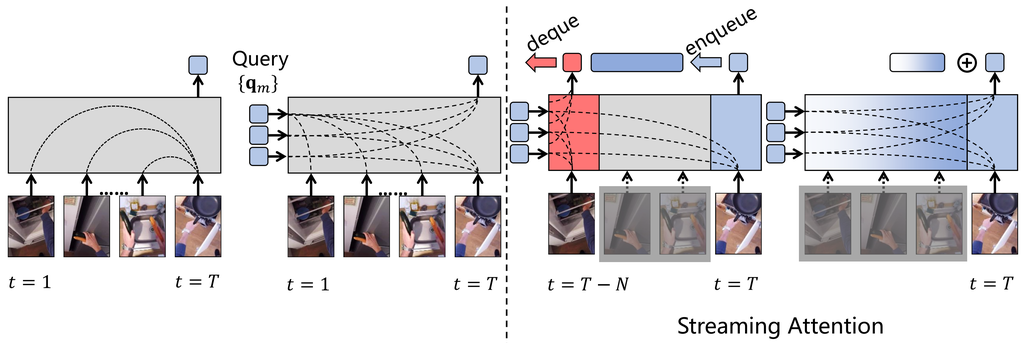
Goal: Streaming video recognition

- It reasons about the action in every frame of a video.

➢ Previous approach: linear/quadratic w/. length

➢ Our approach:

- Reformulate x-attention through the lens of kernel
- Streaming attention with O(1)-update complexity



	Self Attention	Cross Attention	FIFO Attention	Exp. Smoothing Attn.
Runtime	$O(T^2)$	$O(T)$	$O(1)$	$O(1)$
Memory	$O(T^2)$	$O(T)$	$O(N)$	$O(1)$

Attention as kernels:

$$\text{Attention}\left(q_m, x_n \Big|_{n=1}^{t}\right) = \frac{\sum_{n=1}^{t} \kappa(q_m, k_n) v_n}{\sum_{n=1}^{t} \kappa(q_m, k_n)} = \frac{\phi(t)}{\psi(t)}$$

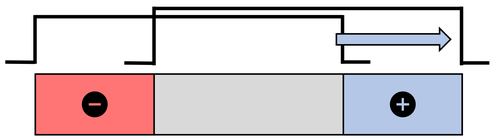
- SoftMax-Attention: $\kappa(q_m, k_n) = \exp\left(\frac{q_m^\top k_n}{\sqrt{C}}\right)$

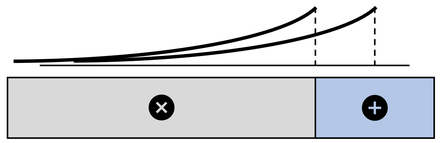
Streaming Attention:

- Decouple temporal and feature component

$$\kappa(q_m, k_n) \to \kappa(\lambda_m + \omega_t, f_n + \omega_n) \to K(\omega_t, \omega_n)\kappa'(\lambda_m, f_n)$$

Box kernel: $K_B(\omega_t, \omega_n) = 1_{[t-n<N]}$ Laplace kernel: $K_L(\omega_t, \omega_n) = e^{-\lambda(t-n)}$

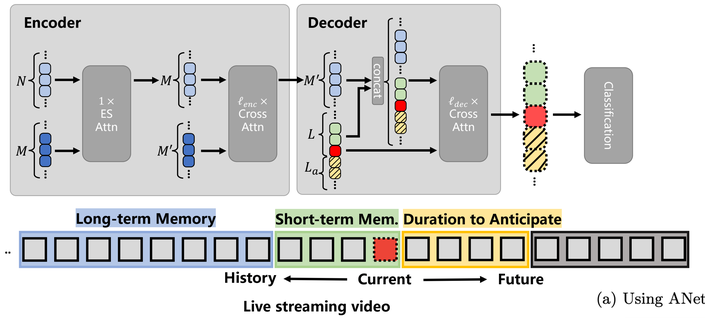




$$\psi(t) = \psi(t-1) + \kappa'(\lambda_m, f_t) - \kappa'(\lambda_m, f_{t-N})$$
$$\phi(t) = \phi(t-1) + \kappa'(\lambda_m, f_t)v_t - \kappa'(\lambda_m, f_{t-N})v_{t-N}$$

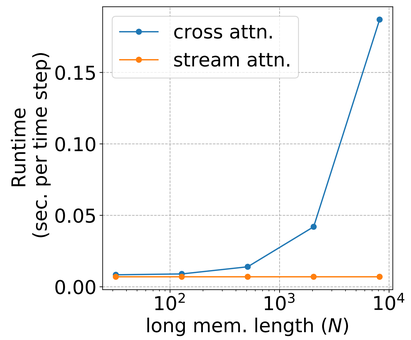
$$\psi(t) = e^{-\lambda}\psi(t-1) + \kappa'(\lambda_m, f_t)$$
$$\phi(t) = e^{-\lambda}\phi(t-1) + \kappa'(\lambda_m, f_t)v_t$$

Overview of TeSTra (Temporal Smoothing Transformer)



Decouple@Long Mem.	Decouple@Short Mem.	mAP on EPIC-KITCHENS
✓	✗	17.0
✗	✗	16.8
✓	✓	15.7

Runtime Comparison



Online action detection on THUMOS14

(a) Using ANet-pretrained feature

Method	mAP
RED [17]	45.3
IDN [13]	50.0
TRN [53]	47.2
OadTR [49]	58.3
LSTR [54]	65.3
Ours	**68.2**

(b) Using Kinetics-pretrained feature

Method	mAP
IDN [13]	60.3
TRN [53]	62.1
OadTR [49]	65.2
LSTR [54]	69.5
Ours‡	67.3
Ours	**71.2**

Online action anticipation on THUMOS14

method	Pre-train	mAP@τ_o 0.25	0.50	0.75	1.0	1.25	1.50	1.75	2.0	average
RED [17]	ANet1.3	45.3	42.1	39.6	37.5	35.8	34.4	33.2	32.1	37.5
TRN [53]		45.1	42.4	40.7	39.1	37.7	36.4	35.3	34.3	38.9
TTM [49]		45.9	43.7	42.4	41.0	39.9	39.4	37.9	37.3	40.9
LSTR [54]		-	-	-	-	-	-	-	-	50.1
Ours		64.7	61.8	58.7	55.7	53.2	51.1	49.2	47.8	**55.3**
TTM [49]	K400	46.8	45.5	44.6	43.6	41.9	41.1	40.4	38.7	42.8
LSTR† [54]		60.4	58.6	56.0	53.3	50.9	48.9	47.1	45.7	52.6
Ours		66.2	63.5	60.5	57.4	54.8	52.6	50.5	48.9	**56.8**