Online Temporal Action Localization with Memory-Augmented Transformer

Youngkil Song* Dongkeun Kim* Minsu Cho Suha Kwak

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

ECCV

Online Temporal Action Localization

Detects actions in a **video stream** and identifies their **the *start*** and **the *end*** times.

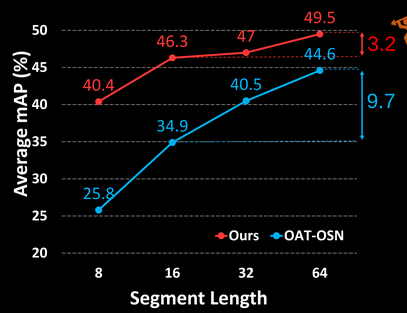
Motivation

- Using only **segment** (short-term memory) is not enough for On-TAL.
- The **memory** utilized is different to predict **the *end*** and **the *start*** of action.

Our Idea

- Use **memory queue** (long-term memory) that selectively stores past information.
- Detect **the *end*** from **segment** and recall **the related *start*** from **memory queue**.

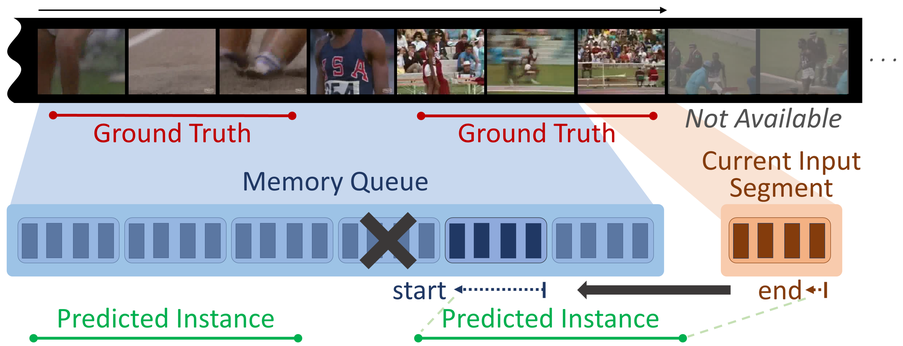
Long-term memory can reduce the *space-time complexity* and improve the *performance* for online temporal action localization!



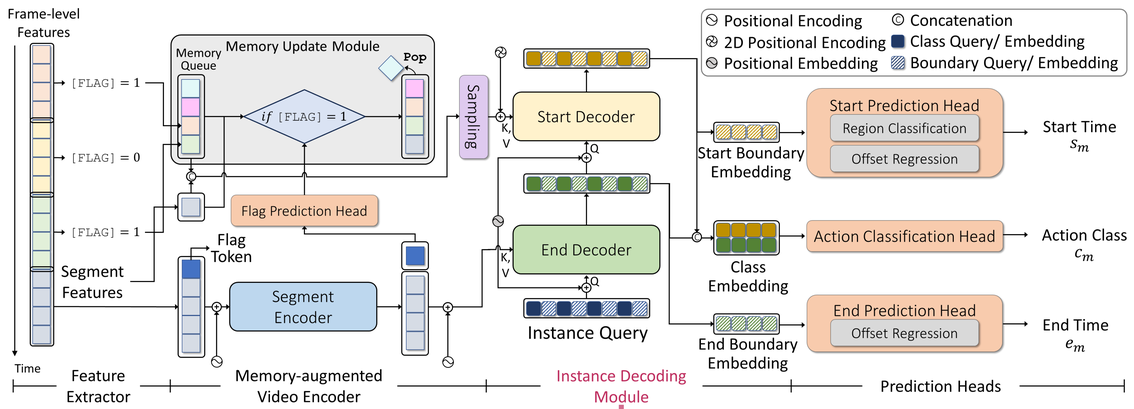
less dependent on the segment size!

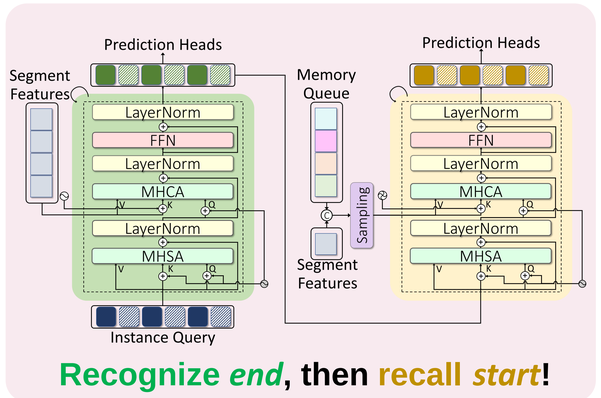
$ can reduce space-time complexity!

Method	Segment size	Inference time	fps	Average mAP
OAT-OSN	64	163.7ms	6.1	44.6
MATR	64	167.1ms	6.0	*49.5*
MATR	*16*	*53.8ms*	*18.6*	46.3



Overall Architecture

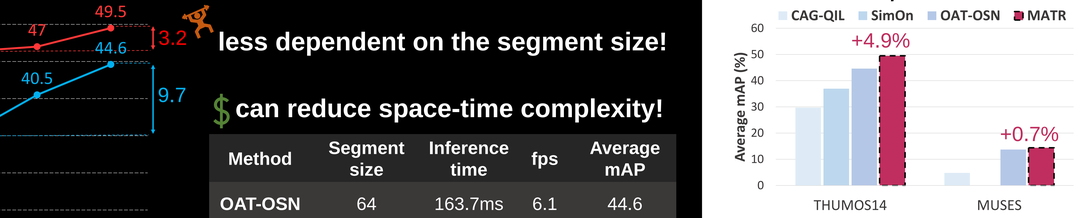




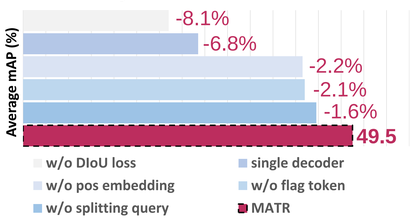
Recognize *end*, then recall *start*!

Experimental Results

Performance comparison



Ablation Study



Impact of memory size

Memory size	0.3	0.4	0.5	0.6	0.7	Average
THUMOS14 dataset						
w/o mem	65.8	58.9	47.6	36.9	20.8	46.0
1	67.2	59.9	50.7	37.9	22.7	47.7
3	67.3	61.9	51.9	37.9	22.6	48.3
7	**70.3**	**62.7**	52.1	38.6	23.7	**49.5**
11	67.9	60.6	49.8	37.8	23.2	47.9
15	66.9	60.1	50.7	38.4	24.1	48.0
19	67.0	60.1	**52.5**	**40.2**	**24.6**	49.1

Memory size	0.3	0.4	0.5	0.6	0.7	Average
MUSES dataset						
w/o mem	22.3	17.6	13.1	9.1	4.6	13.3
1	23.0	18.0	13.5	8.6	4.9	13.6
3	23.1	18.3	13.9	8.9	5.1	13.9
7	**23.5**	18.1	13.5	8.9	4.9	13.8
11	22.9	18.7	14.0	9.4	5.4	14.1
15	**23.5**	**19.3**	**14.3**	9.4	**5.7**	**14.4**
19	22.7	18.9	13.8	**9.6**	5.5	14.1