LVM-Net: Efficient Long-Form Video Reasoning

Saket Gurukar and Asim Kadav

Samsung Research America

Introduction

- Existing video reasoning models operate on few minute videos [1].
- Video reasoning: model's ability to understand three properties: what **activity** is being performed on what **object** over what **time**.

Activity-Query Example:

Q: What activities did I do with {  }

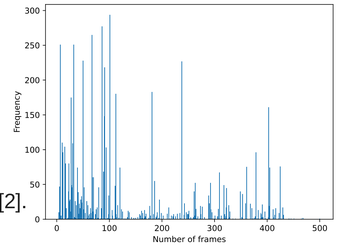
during [00:00 -- 15:00] ?

A: "Making tea"

- Reasoning over hour long videos : challenging on a limited compute budget.
- Existing solutions: use either frame sampling or clip-based aggregation.

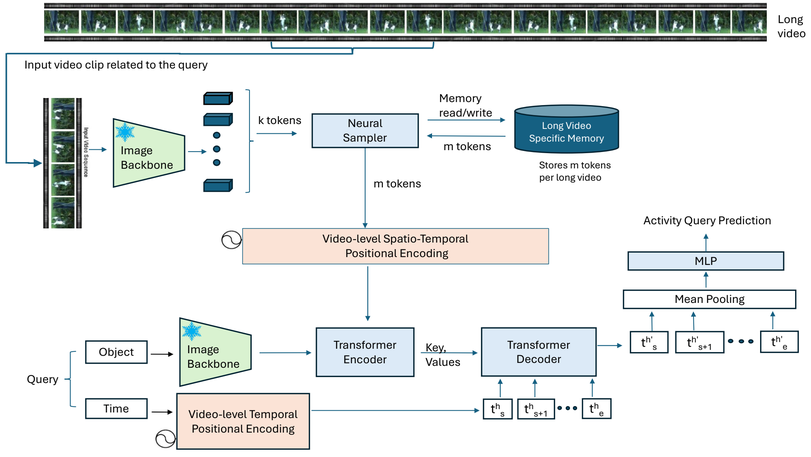
- Another problem: multiple queries over long videos are not supported.
- ReST-ADL dataset has 6000 activity queries on long videos during inference.

- Figure: the number of times "single" frame is reloaded in GPU memory by TubeDETR [2].



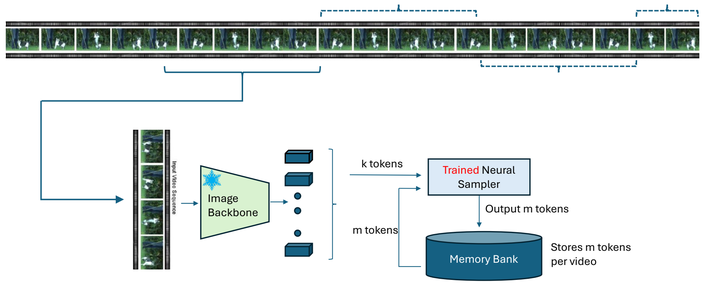
LVM-NET

Performs reasoning over long videos by using attention to store specific information within a fixed memory.

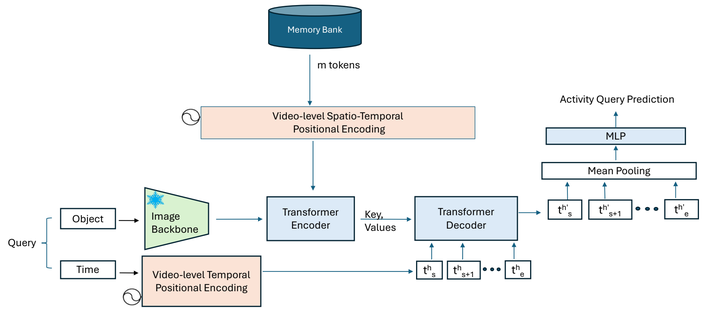


Inference

Two stage inference. Supports multiple queries without reprocessing of frames.



(a) Inference Stage 1



(b) Inference Stage 2

Results

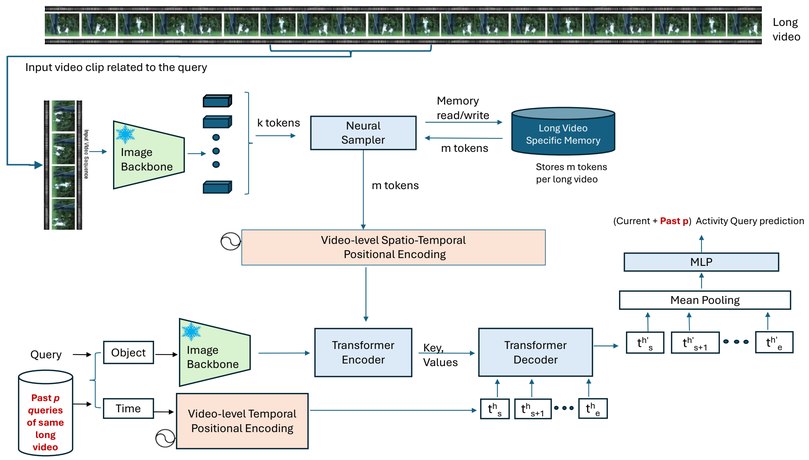
- LVM-Net achieves 18x speedup during inference.
- LVM-Net competitive performance as compared to baselines [1, 2].

Short Queries	
Modified TubeDetr	264 mins
LVM-Net	14 mins (18x speedup)
Medium Queries	
Modified TubeDetr	180 mins
LVM-Net	16 mins (11.2x speedup)
Long Queries	
Modified TubeDetr	174 mins
LVM-Net	15 mins (11.6x speedup)

	Recall@1x	Reject
Short Queries		
ReST system	48.1	68.9
Modified TubeDetr	46.82	74.34
LVM-Net	32.38	68.22
Medium Queries		
ReST system	50.7	63.3
Modified TubeDetr	20.02	63.26
LVM-Net	26.12	71.7
Long Queries		
ReST system	46.3	67.0
Modified TubeDetr	23.11	63.3
LVM-Net	22.8	59.3

Continual Learning

Address sampling bias of neural sampler.



Ablation

- LVM-Net achieves much better performance than random uniform sampling.
- Continual learning loss helps improve the performance

	Recall@1x	Recall@3x
Short Queries		
LVM-Net	32.38	56.78
LVM-Net-random	21.42	43.20
Medium Queries		
LVM-Net	26.12	44.80
LVM-Net-random	18.23	40.57
Long Queries		
LVM-Net	22.81	45.39
LVM-Net-random	18.42	38.45

	Recall@1x	Recall@3x
Short Queries		
LVM-Net	32.38	56.78
LVM-Net-non-continual	26.39	47.31
Medium Queries		
LVM-Net	26.12	44.80
LVM-Net-non-continual	24.81	44.63
Long Queries		
LVM-Net	22.81	45.39
LVM-Net-non-continual	18.28	44.54

References

1. Yang, Xitong, et al. "Relational space-time query in long-form videos." *CVPR*. 2023.
2. Yang, Antoine, et al. "Tubedetr: Spatio-temporal video grounding with transformers." CVPR '22
3. Xie, Sang Michael, and Stefano Ermon. "Reparameterizable subset sampling via continuous relaxations." *IJCAI 2020*