NEC Laboratories America
Relentless passion for innovation

THE STATE UNIVERSITY OF NEW JERSEY RUTGERS

Hopper: Multi-Hop Transformer for Spatiotemporal Reasoning

Honglu Zhou[1]*, Asim Kadav[2], Farley Lai[2], Alexandru Niculescu-Mizil[2], Martin Renqiang Min[2], Mubbasir Kapadia[1], Hans Peter Graf[2]
[1] Department of Computer Science, Rutgers University, Piscataway, NJ, USA. [2] NEC Laboratories America, Inc., San Jose, CA, USA.

GitHub: https://github.com/necla-ml/cater-h

ICLR 2021

Background: Video Reasoning

Existing DL approaches suffer from spatiotemporal biases when applied to video reasoning problems.

Humans subconsciously perform reasoning all the time ...

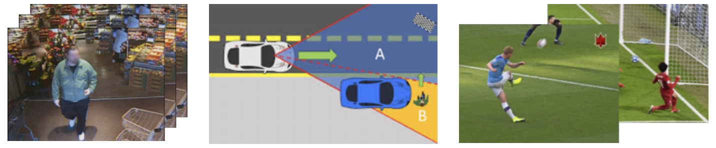
Reasoning matters!

- Human reasoning: Ability to manipulate knowledge entities in terms of relations.
- Machine reasoning: Algebraically manipulate previously acquired knowledge in order to answer a new question.

Task: Object Permanence

Object Permanence: Ability to represent the existence and the trajectory of hidden moving objects.

Learning object permanence requires reasoning!



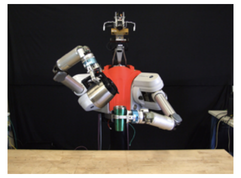
- Shopping: What items the shopper should be billed for?
- Self-Driving: Is there a pedestrian in the front who tries to cross the street?
- Soccer: Which player initiated the pass that resulted in a goal?

Applications

AR/VR

Anti-theft System

Public Safety



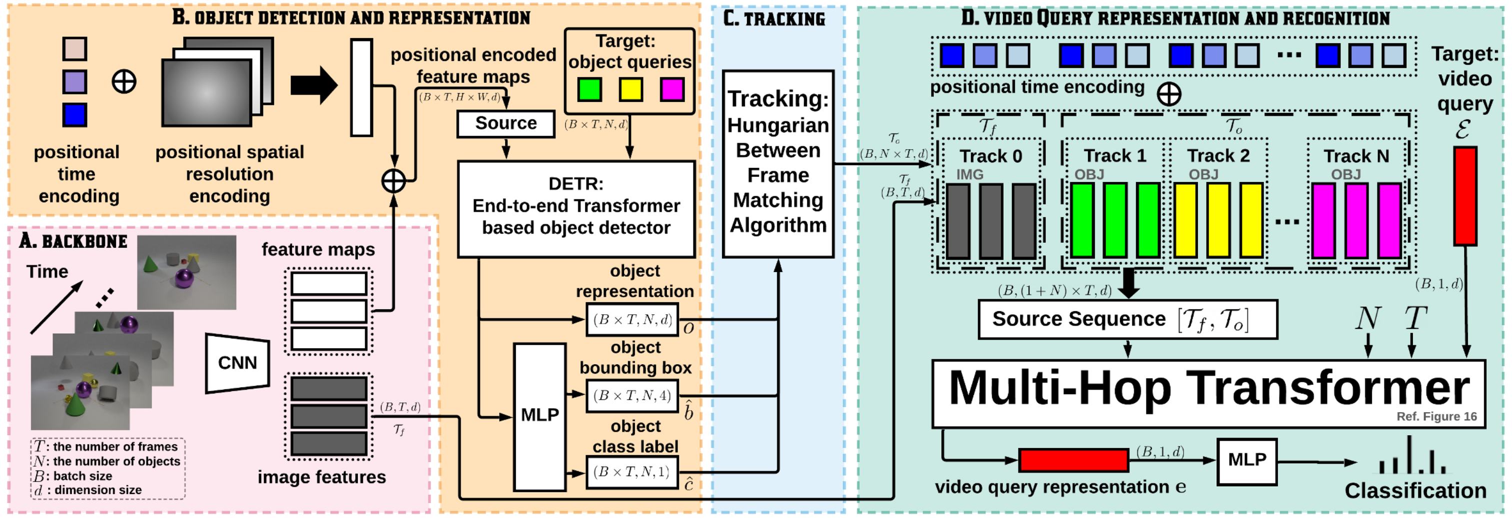
Robotic Manipulation

Retail Intelligence

and so on ...

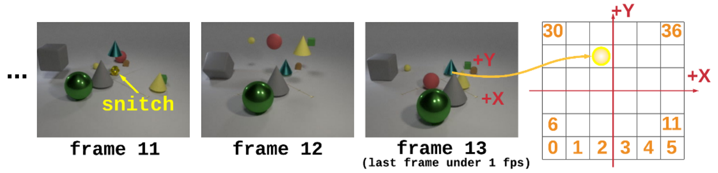
Hopper: A Universal Framework for Video Reasoning

- Object-centric learning: humans think in terms of entities and relations between them.
- Tracking: aggregate sequence features in time order and give consistent feature representations.
- Multi-step compositional reasoning: humans think in steps and understand the world as a sum of its parts.
- Contrastive debiasing: model should not make the correct prediction without seeing the correct evidence.



CATER Dataset

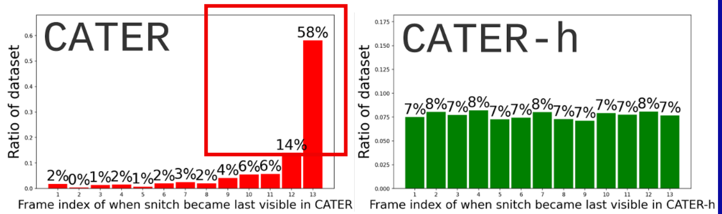
- Snitch Localization under Occlusion and Containment: A diagnostic dataset that implicitly require spatiotemporal understanding and multi-step compositional reasoning.



- Different types of objects & objects move simultaneously.
- Objects can be occluded / contained / carried by other objects
- Every video has a special object called Snitch
- Problem set-up: classification

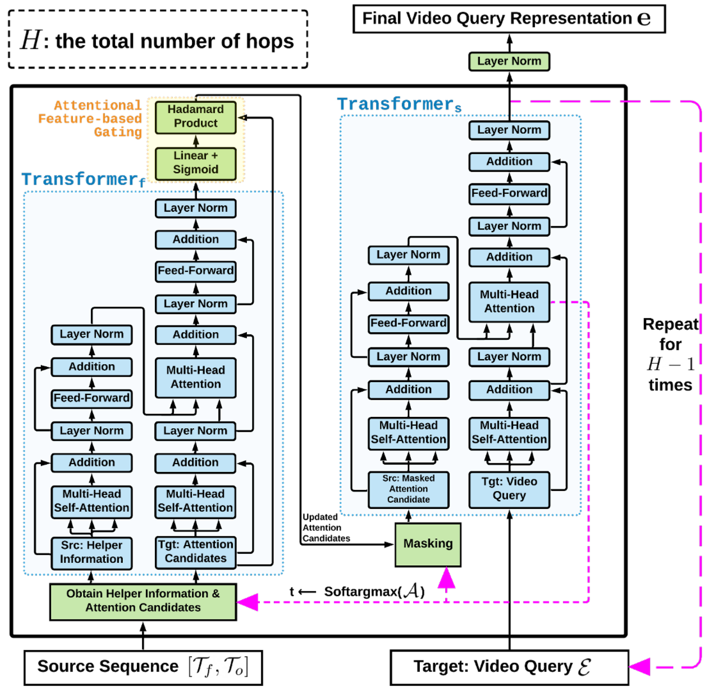
CATER-h: New Unbiased Dataset

- CATER is highly imbalanced in terms of the temporal cues: A diagnostic dataset that implicitly require spatiotemporal understanding and multi-step compositional reasoning.
- CATER-h (CATER-hard): A more difficult video reasoning dataset to avoid any model to achieve high performance by taking shortcut through only looking at the last few frames.



Multi-Hop Transformer (MHT)

MHT reasons by hopping over frames and selectively attending to objects in the frames, until it arrives at the correct object that is the most important for the task.



- Contrastive debiasing loss via masking out:

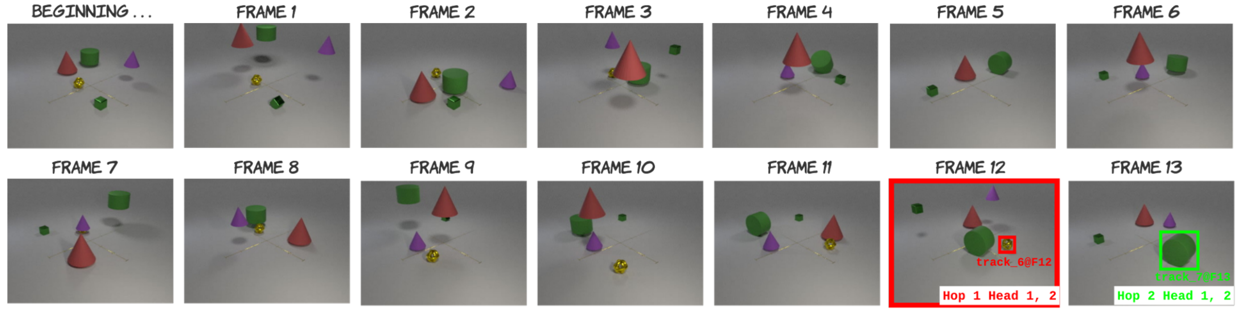
$$\mathcal{L}_{\text{debias}} = \mathbb{E}\left[\sum_{k=1}^{K} g_\theta(\mathcal{M}_{\text{neg}};\cdots)\left(\log g_\theta(\mathcal{M}_{\text{neg}};\cdots)\right)\right]$$

More in the paper!

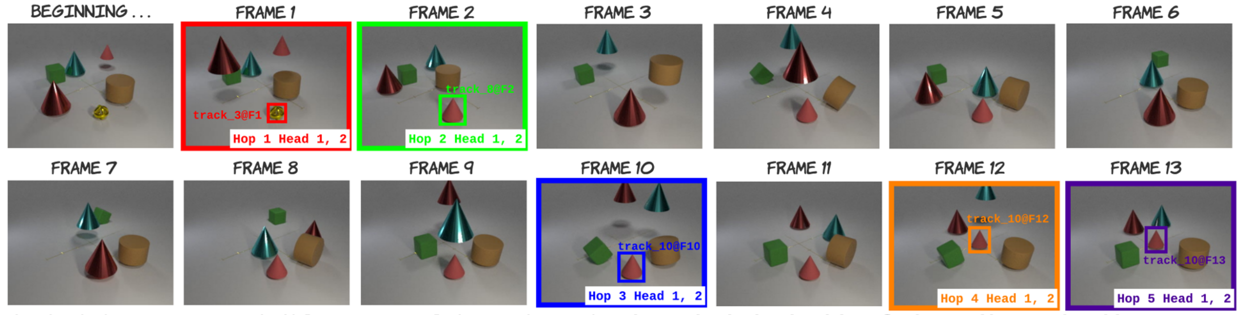
Qualitative Results

Visualizing the most attended object(s) of every hop ...

- MHT provides more transparency to the reasoning process.
- MHT implicitly learns to perform snitch-oriented tracking automatically.



(Occlusion) In the end, *the snitch is occluded by the green cylinder*.



(Snitch becomes not visible very early) In the end, *the snitch is inside of the medium red rubber cone*.

Contributions

- Multi-hop reasoning automatically with interpretability.
- Weak supervision and differentiable reasoning, no supervision for intermediate frames.
- Contrastive debiasing loss to reduce spatiotemporal biases.
- Extensive studies & SOTA accuracy.
- Release a new CATER-h dataset that requires longer reasoning hops.

Potential Impact:
- Improved neural network designs with structural priors encouraging compositional multi-step reasoning.
- Capable for handling any complex time-ordered sequences.

Experiments

Model	CATER		CATER-h	
	Top 1 ↑	L1 ↓	Top 1 ↑	L1 ↓
Random	2.8	3.9	2.4	3.9
DaSiamRPN (Tracking) (Zhu et al., 2018)	33.9	2.4	17.1	2.9
Hungarian (Tracking - ours)	46.0	1.9	37.2	2.3
TPN-101 (Yang et al., 2020)	65.3	1.09	50.2	1.46
TSM-50 (Lin et al., 2019)	64.0	0.93	44.0	1.54
SINet (Ma et al., 2018)	21.1	3.14	18.6	3.24
Transformer (Vaswani et al., 2017)	13.7	3.53	11.6	3.49
Hopper-transformer (last frame)	61.1	1.42	41.8	2.10
Hopper-transformer	64.9	1.11	57.6	1.39
Hopper-sinet	69.1	1.02	62.8	1.25
Hopper-multihop (ours)	***73.2***	***0.85***	***68.4***	***1.09***

Our proposed multi-hop spatiotemporal reasoning model outperforms SOTA.