Hopper: Multi-Hop Transformer for Spatiotemporal Reasoning

Honglu Zhou 1, Asim Kadav 2, Farley La 3, Alexandru Niculescu-Mizil 4, Martin Renqiang Min 5, Muhsin Kapa 1, Hans Peter Graf 1

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

Background: Video Reasoning

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

Task: Object Permanence

- Object Permanence: Ability to understand and remember the presence of hidden objects.

Learning object permanence requires reasoning!

CATER Dataset

- CATER is a challenging dataset that requires reasoning for object permanence.

CATER-h: New Unbiased Dataset

- CATER-h is the new unbiased dataset that has been developed.

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.

Qualitative Results

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