Department of Computer Science
香港城市大學
City University of Hong Kong

COMPUTER SCIENCE COLLOQUIUM

Temporal Learning of Sparse Video-Text Transformers

SPEAKER **Mr Yi Li**

PhD Candidate
Statistical and Visual Computing Lab (SVCL), UC San Diego, USA

DATE 13 Jul, 2023 (Thu)
TIME 1:00 PM - 2:00 PM
VENUE Y6405, 6/F., CS Seminar Room, Yellow Zone, Department of Computer Science, Yeung Kin Man Academic Building, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong

ABSTRACT

Do video-text transformers learn to model temporal relationships across frames? Despite their immense capacity and the abundance of multimodal training data, recent work has revealed the strong tendency of video-text models towards frame-based spatial representations, while temporal reasoning remains largely unsolved. In this work, we identify several key challenges in temporal learning of video-text transformers: the spatiotemporal trade-off from limited network size; the curse of dimensionality for multi-frame modeling; and the diminishing returns of semantic information by extending clip length. Guided by these findings, we propose SViTT, a sparse video-text architecture that performs multi-frame reasoning with significantly lower cost than naive transformers with dense attention. Analogous to graph-based networks, SViTT employs two forms of sparsity: edge sparsity that limits the query-key communications between tokens in self-attention, and node sparsity that discards uninformative visual tokens. Trained with a curriculum which increases model sparsity with the clip length, SViTT outperforms dense transformer baselines on multiple video-text retrieval and question answering benchmarks, with a fraction of computational cost.

BIOGRAPHY

Yi Li is a PhD candidate at the Statistical and Visual Computing Lab (SVCL) at UC San Diego, advised by Professor Nuno Vasconcelos. Prior to joining UCSD, he received the BEng in Electronic Engineering at the Chinese University of Hong Kong in 2017. Yi's research centers around building reliable computer vision systems under various forms of dataset and model bias. Specifically, he works on developing bias mitigation techniques for data and models, applying them to challenging vision problems such as temporal video understanding, and multimodal learning from visual, text and audio modalities.

All are welcome!

In case of questions, please contact Prof Antoni Chan at abchan@cityu.edu.hk, or visit the CS Departmental Seminar Web at https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums.