Towards Compositionality in Video Understanding

Abstract:

Our understanding of the world is naturally hierarchical and structured, and when new concepts are introduced, humans tend to decompose the familiar parts to reason about what they do not know. This leads to the hypothesis that intelligent machines would need to develop a compositional understanding that is robust and generalizable. In this talk, I will discuss our work on compositionality in video understanding from CVPR2022 and NeurIPS2022, as well as a recent preprint, which includes: (i) an object-centric model [1] that directly incorporates object representations into video transformers; (ii) a model [2] that utilizes the structure of a small set of images, whether they are within or outside the domain of interest, available only during training for a video downstream task; (iii) a model [3] that leverages a multi-task prompt learning approach for video transformers, where a shared transformer backbone is enhanced with task-specific prompts. [1] https://arxiv.org/abs/2110.06915 [2] https://arxiv.org/abs/2206.06346 [3] https://arxiv.org/abs/2212.04821 Bio: Roi is a 4th-year CS Ph.D. student at Tel Aviv University and a visiting scholar at Berkeley AI Research Lab (BAIR), working with Prof. Amir Globerson and Prof. Trevor Darrell. Roi is also affiliated as a research scientist at IBM Research AI.