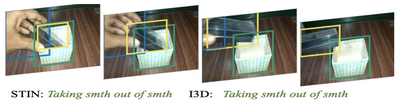
Compositional and Robust Action Understanding

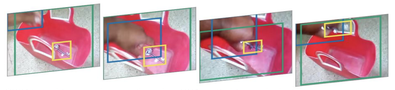
Huijuan Xu huijuan@eecs.berkeley.edu

(1) Compositional action recognition with knowledge reasoning

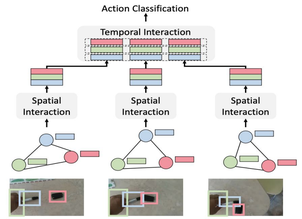
- **Compositional action recognition with spatial-temporal interaction networks. [1]**
  - Study the composition of actions by looking into subject-object interactions.
  - Spatio-Temporal Interaction Networks: explicitly reason about the geometric relations between constituent objects and an agent.



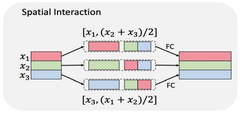
(a) Seen verb and object combination



(b) Unseen verb and object combination

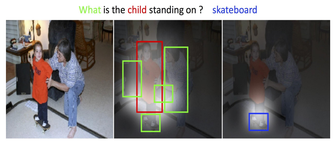


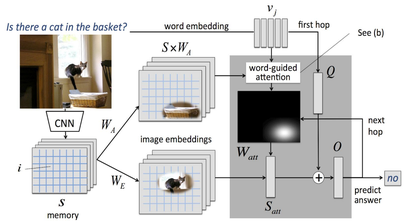
(a) Spatial-Temporal Interaction Network



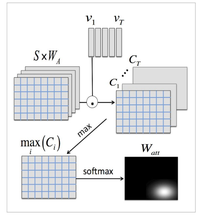
(b) Spatial Interaction Module

- **Attention based reasoning for visual question answering. [2]**
  - Answer a question about a given photograph.
  - Spatial Memory Network VQA (SMem-VQA): incorporate explicit spatial attention into memory networks.





(a) Overview



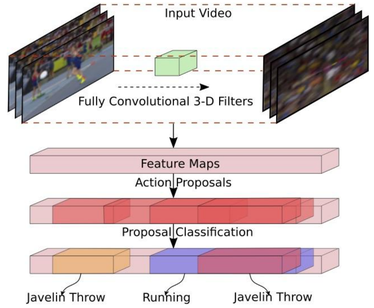
(b) Word-guided attention

- **Object affordance reasoning in compositional action recognition with unseen components.**

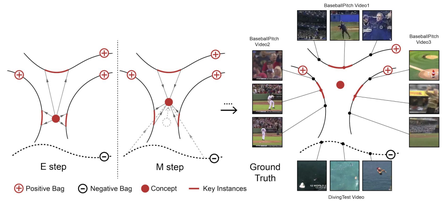
[1]. J. Materzynska, T. Xiao, R. Herzig, H. Xu*, X. Wang*, and T. Darrell*. Something-else: Compositional action recognition with spatial-temporal interaction networks. CVPR2020
[2]. H. Xu, K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. ECCV2016

(2) Multimodal action detection with less supervision

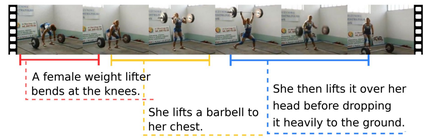
- **Temporal action detection in videos. [3]**
  - Predict the start and end time of activities in untrimmed videos and classify the segments into specific activities.
  - Region Convolutional 3D Network (R-C3D): encode the frame buffer with 3D convolutional filters, propose activity segments, then classify and refine them based on pooled features within their boundaries.



- **Temporal action detection with less supervision.**
  - **Few-shot activity detection. [4]**
  - **Weakly-supervised action localization. [5]**
    - explicitly model the key instance assignment as a hidden variable in Expectation-Maximization (EM) framework following the Multiple Instance Learning (MIL) assumptions.



- **Describe action with language.**
  - **Text-to-clip video retrieval. [6]**
  - **Dense video captioning. [7]**
    - localize distinct events in video stream and generate captions.
    - Joint Event Detection and Description Network (JEDDiNet): design a hierarchical captioning component to model the vision and language context.



[3]. H. Xu, A. Das, K. Saenko. R-C3D: Region convolutional 3d network for temporal activity detection. ICCV2017
[4]. H. Xu, X. Sun, E. Tzeng, A. Das, K. Saenko, and T. Darrell. Revisiting Few-Shot Activity Detection with Class Similarity Control. arXiv preprint, 2020.
[5]. Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu. WeaklySupervised Action Localization with Expectation-Maximization Multi-Instance Learning. ECCV2020
[6]. H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, K. Saenko. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. AAAI2019
[7]. H. Xu, B. Li, V. Ramanishka, L. Sigal, K. Saenko. Joint Event Detection and Description in Continuous Video Streams. WACV2019

In summary, the goal of my research is to build robust action understanding algorithms with human level structural knowledge and multi-modal complementary ability.