Videos as Space-Time Region Graphs

Xiaolong Wang, Abhinav Gupta

In order to recognize a given action in a video sequence like "opening a book", the paper argues that we need two ingredients for this: (1) temporally linking book regions across time while modeling actions as transformations to track how the shape of the book changes across time, and (2) since the state of the objects might change due to some inter human-object and object-object interactions, we also need to model such interactions.

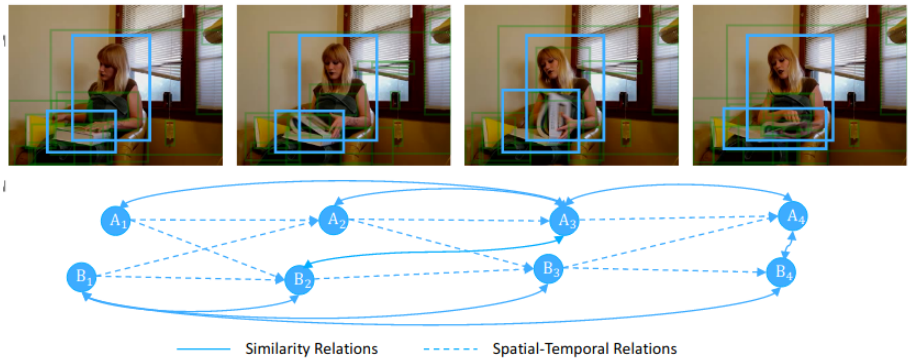

Figure 1. How do you recognize simple actions such as opening book? We argue action understanding requires appearance modeling but also capturing temporal dynamics (how shape of book changes) and functional relationships. We propose to represent videos as space-time region graphs followed by graph convolutions for inference.

To explicitly model such human-object and object-object interactions, the paper proposed to build a space-time graph, where each node in the graph represents the features of a given detected object at a given time step (the features are extracted using ROI Align). The nodes are then connected by two edges, an appearance similarity and spatiotemporal proximity edge. The weights of appearance edges are computed as the similarity to each node to the rest of the nodes (dot product of the transformed features followed by a softmax), while the temporal edges are computed as a weight that reflects the intersection over union between the bounding box of a given object at two different times. The nodes of the graph are then updated using Graph Convolution Networks for both similarity and proximity. The final node representation is computed as a sum of the original features and both the updated similarity and proximity features.

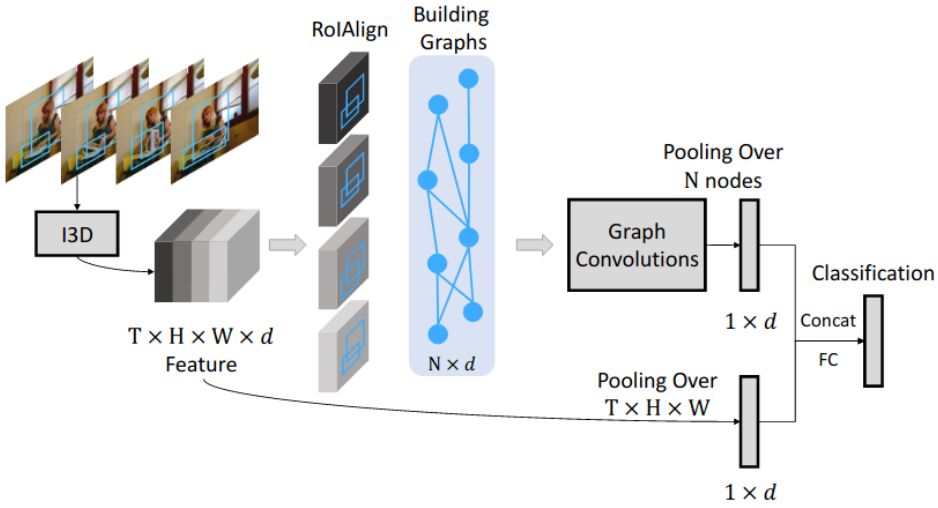

Figure 2. Model Overview. Our model uses 3D convolutions to extract visual features followed by RoIAlign extracting d-dimension feature for each object proposal. These features are provided as inputs to the Graph Convolutional Network which performs information propagation based on spatiotemporal edges. Finally, a d-dimension feature is extracted and appended to another d-dimension video feature to perform classification.