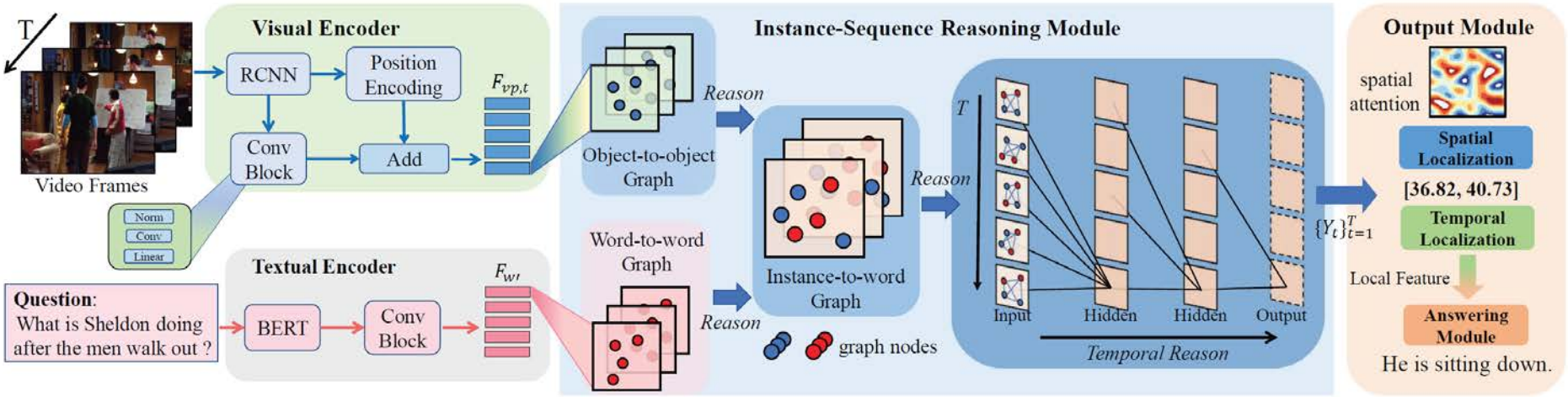
Instance-Sequence Reasoning for Video Question Answering

Rui LIU, Yahong HAN

Frontiers of Computer Science, DOI: 10.1007/s11704-021-1248-1

Problems & Ideas

- Problems of conventional Video QA approaches:
  - Existing methods neglect reasoning from instances to sequence, which is significant for this task.
- Ideas: An Instance-Sequence Reasoning network with instance grounding and temporal localization. Both visual instances and textual representations are embedded into graph nodes.



The overview of our framework. Visual and textual encoders generate corresponding features, $F_{vp,t}$ and $F_{w'}$. In Instance-Sequence Reasoning Module, these features are embedded into intra-modality graph following a single-modal graph "Reason" process. Then, inter-modality graph performs multimodal fusion. In reasoning stage, "Temporal Reason" adopts graph causal convolution (GCC) obtaining final representations Y_t. This Output Module is for TVQA+, including spatial localization module, temporal localization module, and answering module (only for traditional QA).

Main Contributions

- Contributions:
  - A novel Instance-Sequence Reasoning model for Video QA with instance grounding and temporal localization, which uses graph structure to align and fuse visual and textual features;
  - A graph causal convolution on graph sequence, in order to mine the temporal information and reason the question answering. The dynamics of relation between instances and semantic entities are also captured;
  - Superior performance on the benchmarks and related tasks.

Success and failure cases of TVQA+ (Q1, Q2, and Q3), TGIF-QA (Q4), MSVD-QA (Q5) and MSRVTT-QA (Q6).