SEMANTIC-GUIDED OBJECT CLUSTERING FOR MULTI-MODAL REFERRING VIDEO SEGMENTATION

Author's: P. Vinusya[1], S. Rajeshkumar[2], K. Sujith[3], N. Vanjulavalli[4]

Abstract

Referring Video Segmentation (RVS) aims to localize and segment a target object throughout a video sequence based on a natural language description. Unlike traditional video object segmentation, which relies solely on visual cues or predefined object identities, RVS introduces the additional complexity of aligning multi-modal information—namely visual data and textual semantics. This paper presents a Semantic-Assisted Object Clustering framework for Multi-Modal Referring Video Segmentation, designed to enhance object association across temporal frames using integrated semantic guidance.The proposed approach begins by generating object proposals for each video frame, capturing candidate regions potentially corresponding to objects in motion. For each proposal, multi-modal features are extracted, including visual appearance descriptors (color statistics, edge density), motion attributes (spatial position, area, and aspect ratio), and semantic captions derived from lightweight object attribute inference. Simultaneously, the referring expression is encoded using a textual embedding mechanism. A semantic alignment score is computed between the proposal captions and the query embedding, enabling cross-modal similarity estimation.