Image Processing Paper Report

Student ID: 41047902S Student Name: 鄭淮薰

i. Selected Paper

Paper: VLCAP: VISION-LANGUAGE WITH CONTRASTIVE LEARNING FOR COHERENT VIDEO PARAGRAPH CAPTIONING

Conference: IEEE ICIP

Year: 2022

ii. Brief review

VLCAP is a novel approach to generate coherent paragraph descriptions of untrimmed videos using vision-language features and contrastive learning. The proposed approach leverages the human perceiving process, which involves vision and language interaction, to generate a coherent paragraph description of untrimmed videos. The authors propose vision-language (VL) features consisting of two modalities, i.e., (i) vision modality to capture global visual content of the entire scene and (ii) language modality to extract scene elements description of both human and non-human objects (e.g. animals, vehicles, etc), visual and non-visual elements (e.g. actions, attributes).

The proposed contrastive learning VL loss is used for training VLCap. The experiments demonstrate that their proposed VLCap outperforms existing state-of-the-art methods on both ActivityNet Captions and YouCookII datasets.

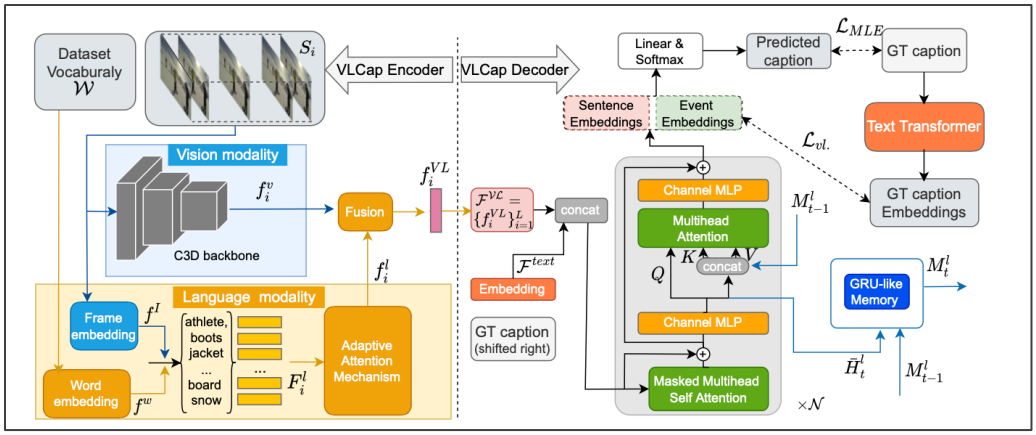

Fig. 1. The flowchart of the VLCAP, as presented in Figure 2 of the VLCAP paper.

The network architecture of VLCAP consists of two parts: the Encoder and the Decoder, which are shown in the figure. The Encoder module is responsible for segmenting the video into individual frames, extracting image features using image processing techniques, and extracting visual and linguistic features from each snippet of the video. To extract visual features, VLCAP uses a pre-trained 3D Convolutional Neural Network (C3D) model in the vision modality. These visual features (f ^{V}) capture the global visual content of the entire scene and serve as the modality of the proposed Visual Language (VL) features. In addition to the global visual content, VLCAP also extracts another visual feature (f I) using a pre-trained Vision Transformer to describe scene elements such as human and non-human objects, visual and non-visual elements in the language modality. The language modality constructs a vocabulary using ground truth captions from the training dataset and encodes each word in the vocabulary into a text feature using a Transformer network. These visual features (f ^{I}) and text features (f ^{W}) are then projected onto the joint embedding space learned by CLIP to obtain image embeddings (I^{e}) and word embeddings (W^{e}) for each snippet. An AAM is then employed to select the most relevant representative language

features (f^{L}) from the visual-linguistic joint embedding space, which serve as another form of VL feature. In the final Decoder stage, a fusion module is used to combine the VL features, which learns to weight each modality according to the importance of generating coherent paragraph descriptions. The fused VL features are then input into a transformer-based decoder, which generates untrimmed video paragraph descriptions.

The applications of this work are in video captioning tasks where generating coherent paragraph descriptions of untrimmed videos is required. This can be useful in various domains such as surveillance, entertainment, education, etc., where automatic video captioning can help in summarizing long videos or providing accessibility for people with hearing impairments.

iii. Comments

Through studying this paper, I understood the entire training and generation process of video captioning, as well as the different types of image processing techniques employed for different purposes. Going deeper into the understanding of each image processing technique used in this paper and their respective objectives was very beneficial for me. For instance, CNN and Transformer are two distinct image-processing techniques used in the VLCAP model for extracting visual and linguistic features from untrimmed videos. CNNs are well-suited for image feature extraction and can capture spatial information from image frames. In contrast, Transformers are better suited for capturing long-term temporal dependencies in language, making them helpful in encoding text features in video captioning tasks. By leveraging the strengths of these two techniques and integrating them into different parts of the model,

VLCAP can extract a comprehensive set of features that capture both global visual content and detailed information about scene elements, as well as language features that describe the relationships between these elements. This results in more accurate and coherent paragraph descriptions of untrimmed videos, which can be useful in various domains such as surveillance, entertainment, and education. However, since VLCAP incorporates multiple methods and modules, I think the training may be computationally demanding. One possible solution could be reducing certain modules or replacing them with lighter ones to improve efficiency while sacrificing some precision.