모멘트 검색 및 하이라이트 탐지를 위한 태스크 맞춤형 인코더-디코더 프레임워크

Task-Tailored Encoder-Decoder Framework for Moment Retrieval and Highlight Detection

Jongseong Bae, Kisu Lee, Won-Yong Shin, Ha Young Kim*
Yonsei University

{js.bae, kisu0928, wy.shin, hayoung.kim}@yonsei.ac.kr

Abstract

In recent years, moment retrieval and highlight detection (MR/HD) have garnered significant attention as major tasks in video understanding. Due to the similarity between these tasks, they are often jointly addressed using a single unified model architecture in previous works while overlooking the specific characteristics of each task. To elaborate, MR aims to predict moment spans related to a query, while HD focuses on predicting discrete saliency scores for each clip. To address this issue, we propose a simple yet effective encoder-decoder framework for MR/HD, named T2QD-DETR. This framework separates the intermediate representation learning for each task to train task-tailored features. T2QD-DETR learns discrete clip-wise representations from the multi-modal encoder using the objective tailored for HD. For MR, the model emphasizes temporal information in the clip-wise features by employing a separate decoder while keeping the encoding part unaffected. Through experiments conducted on an MR and HD benchmark dataset, we demonstrate the superiority of the proposed T2QD-DETR over the robust baseline model.

Ⅰ. Introduction

As numerous videos are widely being spread on various online video platforms, moment retrieval and highlight detection (MR/HD) for video understanding have been in the spotlight recently. Given a pair of a video and a text query, MR and HD explore related parts in the video to the query. To elaborate, MR focuses on predicting related continuous spans in the input video, and HD aims to predict the saliency scores for each clip in the video.

In previous works [1,2], MR and HD have been jointly addressed through a single architecture with two heads for each task, due to the similarity in their objectives. Moment-DETR [1] and QD-DETR [2] fuse multi-modality features (video and text) utilizing transformer-based encoders and then predict saliency scores using simple predictors and moment spans using transformer decoders. Thus, the feature fusion encoders are jointly trained for both tasks.

Despite the efficiency of the current joint learning paradigm, it overlooks these individual characteristics, limiting the model's ability to capture task-specific feature patterns. MR requires a temporal understanding of the input video since it aims to predict moment spans. On the other hand, HD necessitates understanding discretized clip-level relationships between modalities due to predicting independent scores per video clip.

To handle this issue, we introduce a novel task-tailored training framework, T2QD-DETR, which extends the architecture of QD-DETR. T2QD-DETR trains the feature fusion encoder to extract the well-discretized relationships between modalities, focusing on the objective of HD since the encoder outputs clip-wise fused features. Simultaneously, the encoder outputs are fed to a decoder for MR with a recurrent neural network (RNN), which aims to emphasize temporal information in the video. T2QD-DETR trains the decoder separately from the encoder using the stop-gradient technique to prevent disturbing the encoder that learns clip-wise relationships. In this manner, the encoder of our T2QD-DETR concentrates on learning the clip-level representations for HD, and based on them, the decoder is specialized to capture the inter-clip relationship, which is fundamental information for MR. Our experimental results on a MR/HD benchmark, QVHighlights [1], support the effectiveness of the proposed T2QD-DETR.

Ⅱ. Method

Given a video of M clips and a text query of N words, the purpose of MR is to predict moment spans related to the text with a format of the center coordinates and width, while HD aims to predict saliency scores per clip. Following previous works [1,2], we extract video and text features, $X_v \in R^{M \times C_v}$ and $X_t \in R^{N \times C_t}$, where C_v and C_t denote the output dimensions, using frozen video and text encoders, respectively. T2QD-DETR projects each feature into a shared d dimensional embedding space using sibling multi-layer perceptrons, following [1]. Thus, the video and text features are projected to $X'_v \in R^{M \times d}$ and $X'_t \in R^{N \times d}$, respectively. Subsequentially, T2QD-DETR extracts query-dependent video features $X_{enc} \in R^{M \times d}$ from cross- and self-attention-based transformer encoders, inspired by [2]. X_{enc} is fed into MR and HD prediction heads, the transformer decoder and a linear projection, respectively, and each prediction head outputs final predictions for MR and HD.

In traditional QD-DETR [2], the encoders are jointly learned across the MR and HD heads. However, MR necessitates predicting the continuous moment spans requiring temporal information in video, while HD requires clip-level independent predictions on discretized representation. We assume that the differences in the characteristics of the objectives for MR and HD may cause confusion during the joint learning of the encoders. Meanwhile, the query-dependent features have a discrete characteristic since they present clip-wise representations. Thus, T2QD-DETR trains the encoders only with the HD head by applying a stop-gradient to the MR head. Furthermore, we add an RNN-based layer to the transformer decoder to enhance the MR branch's ability to reflect the inter-clip relational information. In this way, T2QD-DETR boosts both MR and HD performances by helping the model learn task-tailored intermediate representations.

Ⅲ. Experimental Results

To evaluate the MR/HD performances of proposed T2QD-DETR, we utilize QVHighlights validation split [1]. For evaluation metrics, following those in the baseline, we report mean average precision (mAP) for MR, and mAP and HIT@1 score for HD.

Table 1. Experimental result

	MR	HD	
	Avg. mAP	mAP	HIT@1
QD-DETR	39.86	38.94	62.40
+ stop-gradient	40.08	39.89	63.48
T2QD-DETR	**43.27**	**39.97**	**64.58**

Table 1 shows the MR/HD performance comparison of QD-DETR and T2QD-DETR, as supporting the effectiveness of proposed method. By simply applying stop-gradient to MR branch of X_{enc}, we can observe the performance improvements in both tasks.

In addition, T2QD-DETR adds a long short-term memory layer to MR's decoder for enhancing temporal information. T2QD-

*Corresponding author.

DETR achieves impressive performance enhancements compared to QD-DETR. This result explains the significance of task-specialized feature learning in MR and HD. Furthermore, we infer that temporal information from videos has a meaningful role for MR.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2023R1A2C200337911 and No. RS-2023-00220762).

REFERENCES

[1] Lei, J., Berg, T.L. and Bansal, M., 2021. Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems, 34, pp.11846-11858.

[2] Moon, W., Hyun, S., Park, S., Park, D., & Heo, J. P. (2023). Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 23023-23033).