Highlight Prediction of Esports Videos Using Multimodal Transformer Models

Assel Islam[1], Yoonji Ryu[2], Wonseok Jang[2], Hyunwoong Pyun[2], Gyemin Lee[1*]

***Abstract*— In this paper, a novel highlight prediction method for esports videos is proposed. We investigate Transformer to understand contexts of long-length videos. Our model combines frame-level contexts and shot-level contexts. Moreover, our model uses both visual and audio features. The proposed model is evaluated on esports videos of League of Legends Champions Korea matches. Experiment results suggest that the proposed method produces quality esports highlights.**

I. Introduction

Esports is one of the fastest-growing sports industries in the world. Unlike traditional sports, esports games are mostly distributed through online streaming platforms. As the media consumption pattern is shifting in recent years, highlight video is emerging as a popular form to consume esports games. In this regard, the demand for automatic highlight prediction is increasing to facilitate efficient distribution and to attract new viewers. In this paper, we propose a novel highlight prediction method for esports videos.

II. Method

In predicting highlight of long videos, capturing contexts over longer period of time is critical. For example, a certain move of a player will be proven to be important if it leads to more scores later in the game.

In this paper, we investigate Transformer to extract long-term contexts from esports videos. In addition, we use both visual and audio features to improve highlight prediction.

While Transformer is known to be more effective than RNNs in handling long sequence data, highlight prediction of very long videos still remains challenging. Thus, we propose to combine both the frame-level and the shot-level contexts using the attention mechanism of Transformer. We also merge visual and audio features twice, once in frame-level and once in shot-level. As illustrated in Fig. 1, the multimodal shot-level contexts encoded with Transformer are cross-attended with frame-level contexts using Transformer Decoders. These multi-level contexts are concatenated to predict the importance scores with a Fully-Connected layer.

III. Experiments and Results

Dataset. The proposed method is demonstrated on an esports dataset. The dataset consists of 64 matches of League of Legends Champions Korea held in 2017. Each match varies from 25 to 80 minutes in length, 35 minutes on average. Editorial highlight videos are accompanied with full videos and treated as ground truth. We used 57 matches for training and seven for testing.

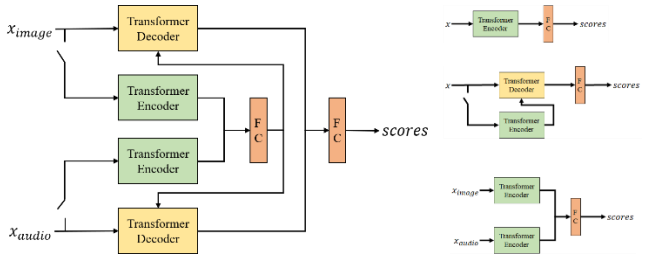

Figure 1. The proposed model combines visual and audio features to predict hightlight of esports videos. The Transformer layers help to capture long-term contexts. Baseline models are on the right.

Implementation details. Visual features are extracted using ResNet-34 pretrained on ImageNet after subsampling videos by 1 fps. For audio, 500-dim MFCC features are extracted each second. Audio features are projected to 512-dim, the size of visual features. The number of heads of all Transformers is set to 16. The learning rate is 1e-4 and the weight decay is 1e-3. Models are evaluated after training 20 epochs. As the output of the model is importance scores, we select video segments of top 10% scores to predict highlight.

TABLE I. Results of various Transformer-based hilight prediction models on League of Legends esports dataset.

Data type	Method	F-score (%)
Image	Encoder	58.13
	Encoder+Decoder	58.26
Audio	Encoder	59.18
	Encoder+Decoder	58.93
Image + Audio	Multi-Encoder	65.99
	Proposed	**67.18**

Results. The proposed method is evaluated with F-scores. Our method is compared with baseline Transformer models. As can be seen in Table I, our method performs the best. The multimodal models that use both image and audio features outperforms the models that use only one type of features. Furthermore, the proposed method improves the simple multimodal model by more effectively combining the frame-level and the shot-level contexts from multimodal features.

*Research supported by National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1I1A4064440).

[1]The Department of Smart ICT Convergence Engineering, Seoul National University of Science and Technology, Seoul, Korea.

[2]College of Sport Science, Sungkyunkwan University, Suwon, Korea.
e-mail: [1]{assel, gyemin}@seoultech.ac.kr,
[2]{blessryu, wjang, hwpyun}@skku.edu