Highlight Moment Detection System

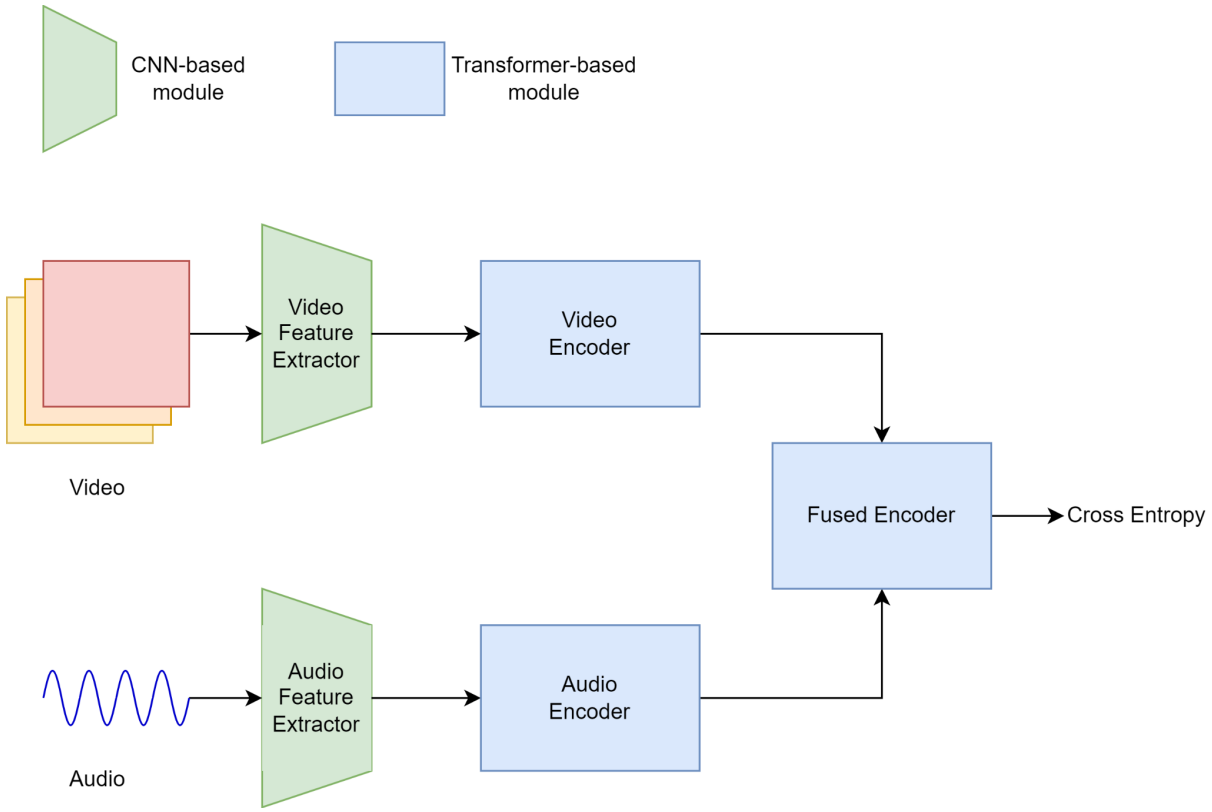
1. Problem definition and scope

The goal of this task is to automatically identify the most important and highlight moments in FIFA 23. As with any supervised learning task, it requires a labeled dataset of videos, which are sequences of images, and their corresponding moments for training. The output of the model is a set of highlight segments or timestamps.

2. Approach

Highlight moments in gaming refer to certain events or actions that are particularly exciting or noteworthy, such as scoring a goal in a sports game or achieving a difficult feat in a video game. These moments often elicit strong emotional reactions from players, such as cheers, screams, or other expressions of excitement. To effectively detect and classify these highlight moments, a multi-modal approach is needed.

A multi-modal approach involves using multiple sources of data, such as both video and audio, to make predictions or decisions. In the case of highlight moment detection in gaming, the main idea is to separately encode the video and audio sequences using different sub-networks, and then combine the encoded representations before passing them through a joint network. This allows the model to take advantage of both visual and auditory information to make more accurate predictions.



To train the model, a common objective function used in classification tasks is Cross Entropy. Cross Entropy measures the dissimilarity between the predicted and actual class labels, and aims to minimize this dissimilarity during training. By using Cross Entropy as the objective function, the model can learn to distinguish between different types of highlight moments and accurately classify them.

Due to the large amount of information contained in both the video and audio data, a Video Feature Extractor and an Audio Feature Extractor are used to sample and extract relevant features from the data, respectively. These modules are based on Convolutional Neural Networks (CNNs), which are commonly used for image and audio processing tasks. The video frames could be recorded in 60FPS or higher, while only a few frames per second may be needed to capture enough information. After being sampled, video frames are extracted into compact representation. Similarly, the audio is usually recorded at a sample rate of 44kHz, it should be transformed into a spectrogram (an informative 2D data form of speech) before being sampled and extracted.

The Transformer-based architecture has become widely used for representing time-series and sequence-styled data. To represent the video and audio data for identifying highlight moments in FIFA 23, I suggest using the Transformer[1] and Conformer[2] (a variant of Transformer with CNN layers), respectively. The Transformer is particularly effective at representing global relationships between data, which is important for video data because frames are dependent on each other. Similarly, audio data can also be represented globally with the Transformer. However, because audio data also has local relationships, the Conformer is chosen to capture these relationships. After representing in separate Encoder, the feature maps of video and audio are combined and forwarded through a fused Transformer network.

To classify highlight moments in frames, my approach is to employ Cross Entropy as the objective function. The sequence labels are divided into segments that are assigned to one of two classes - either highlight or common frames. The feature maps are then input into the Softmax function to determine the output probabilities, which are subsequently passed into the loss function.

3. Metrics

The standard evaluation metrics, including Precision, Recall, and F1 score, are utilized to assess the performance of the model.

4. Challenges

The most difficult aspect of the task is the serving part, which requires the model to operate in real-time. The model must accurately detect the start and end frames of highlight moments to begin and end recording, respectively, which means it cannot rely on future frames.

One potential solution to this problem is to train the model in a chunk-wise and cache-wise manner. The videos and audios are divided into segments or "chunks". During each iteration of the training process, a new chunk is added to the cache and passed through the network. In the inference phase, the cache is also passed through the network, but only the last chunk is used for the prediction. By implementing this approach, the model can better handle real-time serving and make accurate predictions for highlight moments.

---

[1] Vaswani et al., “Attention Is All You Need.”
[2] Gulati et al., “Conformer.”