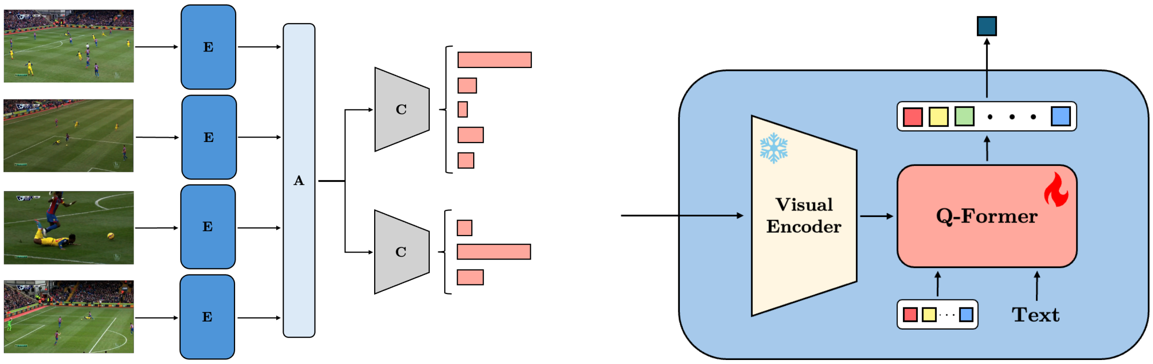
VF-VARS: Leveraging Video Foundation Models for Video Assistant Referee Systems

Jihwan Hong[*] **Youngseo Kim**[*] **Junseok Lee** **Jimin Lee** **Sehyung Kim** **Hyunwoo J. Kim**[†]

Department of Computer Science and Engineering, Korea University
{csjihwanh, xwsa568, behindstarter42, 2001joe, shkim129, hyunwoojkim}@korea.ac.kr



(a) Overall architecture

(b) Encoder module

Figure 1. (Left) Overall architecture: E represents the encoder module, A denotes the aggregation, and C indicates the classification head. (Right) Encoder module consisting of the Vision encoder and Q-Former.

1. Introduction

The existing VARS model [1] extracts video features from multiple clips and uses attention mechanisms to aggregate these features for classification. However, this approach presents two key challenges. Firstly, the number of views varies between 2 and 4, leading to a training process that involves randomly selecting only two views. This results in suboptimal outcomes because the number of views during testing is not always two. Secondly, the mViT [4] model employs temporal pooling, which may not be optimal for foul recognition tasks where specific frames are more important than others. To address these issues, we have designed a framework that handles the varying number of views and leverages a Video Foundation Model (VFM) to encode the video before classification, aiming to improve consistency and accuracy. Our code is available at https://github.com/Jordano-Jackson/soccernet-MLV.

2. Methodology

2.1. Preprocessing

During the dataset preprocessing stage, we encountered video clips with varying numbers of views, precisely 2, 3, or 4 views. To standardize the number of views, we adjusted the views accordingly. For video clips with two views, the second was duplicated twice to create four views. This was because the first view always showed the overall scene, while views 2, 3, and 4 were primarily close-up views. Therefore, to maintain consistency, we duplicated only the second view. For video clips with three views, one view was randomly selected from the second or third view and duplicated to achieve four views in total. This method ensured that all video clips had the same number of views, optimizing the performance of the attention mechanism. Additionally, different augmentations were applied to each view to make them distinct, enhancing the model's ability to generalize from the data.

2.2. Architecture

Our approach primarily leverages the capabilities of VideoChat2 [3]. Additionally, our structure largely retains

[*] Equal contribution.
[†] Corresponding author.

the architecture of VARS [1]. We extracted video features from the clips by utilizing the vision encoder and subsequent Q-Former from VideoChat2. These features, obtained from each view, were then aggregated using an attention mechanism similar to the VARS framework.

We kept the visual encoder frozen during training to prevent catastrophic forgetting and reduce training costs. We then fine-tuned the Q-Former, including its learnable query inputs. Additionally, we provided the Q-Former with textual descriptions of each foul and the degrees of severity as instructions [1, 2].

The Q-Former outputs 96 tokens, each with 768 dimensions, which were average pooled. After enhancing these features with a 3-layer MLP, we added a classification head for severity and action classifications. We utilized a weighted cross-entropy loss, identical to the one used in VARS. The model was trained for approximately 50 epochs with a 2×10^{-4} learning rate.

References

[1] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. VARS: Video assistant referee system for automated soccer decision making from multiple views. pages 5085–5096, 2023. 1, 2

[2] IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, 2022. 2

[3] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023. 1

[4] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1