Video Captioning for Soccer Games: Evaluation of Pretrained Models and Training Strategies

Haki Ali Qasim Saeed
Khadija Mehmood
fast.nu@edu.com

***Abstract*—Video captioning is an essential task in multimodal learning, especially for domain-specific scenarios such as soccer games. This paper evaluates multiple pretrained models, including GPT-2, BERT, VideoMAE-base, and a VideoCLIP model, for generating captions from soccer videos. We analyze training strategies, such as fine-tuning the entire network and freezing encoder layers, to improve computational efficiency and caption coherence. The highest BLEU score of 0.51 was achieved using GPT-2 with frozen encoder layers, producing semantically rich captions. Comparatively, VideoCLIP achieved a BLEU score of 0.36, and BERT + VideoMAE-base achieved 0.32. This paper highlights the trade-offs between training complexity, computational cost, and caption quality.**

***Index Terms*—Video captioning, GPT-2, BERT, VideoMAE-base, VideoCLIP, multimodal learning, BLEU score.**

I. Introduction

Video captioning bridges the gap between computer vision and natural language processing by generating textual descriptions from visual data. Soccer game videos provide unique challenges due to their dynamic nature, complex interactions, and domain-specific terminologies. Effective video captioning in this context can enable automated highlight generation, assistive commentary, and enhanced user engagement.

This study investigates the application of pretrained models for captioning soccer videos. Specifically, we evaluate the performance of GPT-2 with a transformer decoder, BERT paired with VideoMAE-base, and a VideoCLIP model. The primary contributions of this work are:

- Employing and evaluating multiple pretrained architectures for domain-specific video captioning.
- Analyzing the impact of freezing encoder layers on computational efficiency and BLEU scores.
- Providing a comparative analysis of caption quality and training complexity across models.

II. Related Work

A. Video Captioning

Video captioning involves generating natural language descriptions of video content. Traditional approaches use sequence-to-sequence models, while more recent methods leverage transformers and multimodal learning.

B. Pretrained Models

Pretrained models, such as GPT-2 and BERT, have demonstrated success in text generation tasks. Video-specific models like VideoMAE-base and VideoCLIP excel in extracting spatiotemporal features for downstream tasks.

III. Dataset and Preprocessing

We utilized a curated dataset of soccer videos, annotated with textual descriptions of key events. The dataset includes diverse scenarios such as goals, fouls, and passes. Preprocessing steps involved:

- Frame extraction using pretrained vision encoders.
- Alignment of frames with corresponding textual annotations.
- Normalization and resizing of video frames to 224×224.

IV. Methodology

A. Model Architectures

1) GPT-2 with Transformer Decoder: The model integrates a pretrained GPT-2 decoder with a video encoder, fine-tuned for soccer-specific captions.

2) BERT + VideoMAE-base: A two-stage pipeline where VideoMAE-base extracts spatiotemporal features, and BERT generates captions.

3) VideoCLIP: VideoCLIP leverages contrastive learning for joint video-text representations, enabling efficient caption generation.

B. Training Strategies

We experimented with two training strategies:

- **Fine-Tuning Entire Network:** Training all layers of the encoder and decoder.
- **Freezing Encoder Layers:** Keeping encoder layers fixed while fine-tuning the decoder.

C. Evaluation Metrics

Performance was measured using BLEU scores and qualitative assessments of caption coherence.

V. Results and Discussion

A. Quantitative Results

Table I summarizes the BLEU scores and training times for each model.

TABLE I
Model Performance and Training Time

Model	BLEU Score	Training Time	Observations
GPT-2 (frozen encoder)	0.51	50 hrs	Best coherence
VideoCLIP	0.36	25 hrs	Moderate quality
BERT + VideoMAE-base	0.32	15 hrs	Fastest training

B. Qualitative Results

Examples of generated captions:

- **GPT-2:** "The striker scores a goal with a powerful shot."
- **VideoCLIP:** "Player runs and kicks ball to score."
- **BERT + VideoMAE-base:** "Goal shot ball kick."

C. Analysis

Freezing encoder layers improved GPT-2's BLEU score and reduced overfitting. However, it required significant computational resources. VideoCLIP achieved a balance between quality and efficiency, while BERT + VideoMAE-base prioritized speed over coherence.

VI. Conclusion

This paper demonstrates the potential of pretrained models for soccer video captioning. While GPT-2 achieved the highest BLEU score, VideoCLIP provided a balanced trade-off between efficiency and quality. Future work includes experimenting with larger datasets and exploring multimodal transformers.

Acknowledgment

We thank [Funding Agency/Institution] for supporting this research.

References