Laboratory File

on

AGENTIC AI

School of Engineering and Technology

Department of Computer Science and Engineering

Subject code – **CSCR 3215**

SUBMITTED BY:

Name: Mohd Zaid
System ID: 2023307325

SUBMITTED TO:

Mr. Ayush Singh

Sharda University

Greater Noida, Uttar Pradesh

Lab 01: Fine-Tuning
Project Documentation: Football Image Captioning Using BLIP

1. Project Objective

The primary goal of this project is to adapt a pre-trained BLIP (Bootstrapped Language-Image Pretraining) model to automatically generate accurate, context-aware captions for football-related images. By fine-tuning the model on domain-specific data, the system learns to understand both visual elements and their semantic relationships, resulting in more meaningful image descriptions.

2. Methodology

2.1 Dataset Acquisition
A football-specific image caption dataset was obtained from Hugging Face, consisting of football images paired with descriptive textual captions. This dataset provides the necessary multimodal data for supervised fine-tuning.

2.2 Data Preprocessing
- Images and captions are processed using the AutoProcessor from the Transformers library.
- Images are transformed into pixel-level embeddings suitable for the vision encoder.
- Text captions are tokenized into numerical representations.
- A custom PyTorch Dataset and DataLoader are implemented to support efficient batching and training.

2.3 Model Architecture
The BLIP Image Captioning Base model is utilized, which combines:
- A vision encoder for extracting meaningful visual features from images.
- A text decoder that generates natural language captions based on these features.

This architecture enables effective learning across both visual and textual modalities.

2.4 Model Training
- The model is fine-tuned using the AdamW optimizer.
- Caption generation is trained using cross-entropy loss, with caption tokens serving as ground-truth labels.
- Backpropagation is performed over multiple epochs to minimize prediction error and improve caption quality.

3. System Working

1. The input football image is fed into the vision encoder, which extracts high-level visual representations.
2. These visual features are passed to the text decoder.
3. During training, the decoder learns to associate visual features with the correct sequence of caption tokens.
4. During inference, the model generates captions token-by-token, relying only on the image input.
5. The predicted token IDs are decoded to produce a human-readable caption.

4. Results and Outcomes

- The fine-tuned model generates accurate and football-specific captions.
- Caption quality improves significantly compared to the base pre-trained model.
- Domain-specific training enhances contextual understanding (players, actions, stadiums, etc.).
- The trained model is successfully uploaded to the Hugging Face Hub for reuse and deployment.

- The project validates the effectiveness of multimodal transformer-based architectures.

5. Conclusion

This project demonstrates that fine-tuning vision-language transformer models such as BLIP can substantially enhance image captioning performance in specialized domains. The approach effectively bridges visual perception and language generation. Furthermore, the system is scalable and can be extended to other application areas, including medical imaging, wildlife monitoring, and surveillance systems.