

**Laboratory File**  
**on**  
**AGENTIC AI**



School of Engineering and Technology  
Department of Computer Science and Engineering  
**Subject Code – CSCR 3215**

**SUBMITTED BY:**

**Name: Mohd Suhail Khan**  
**System ID: 2023418892**

**SUBMITTED TO:**

**Mr. Ayush Singh**

**Sharda University**  
**Greater Noida, Uttar Pradesh**

## Lab 01: Fine-Tuning

### Finetune BLIP On An Image Captioning Dataset

#### Objective:

The aim of this project is to fine-tune a pre-trained **BLIP (Bootstrapped Language–Image Pretraining)** model to automatically generate accurate and meaningful captions for football images by jointly understanding visual and textual information.

#### Methodology:

1. **Dataset Collection:** A football image–caption dataset was sourced from Hugging Face, containing images paired with descriptive textual captions.
2. **Data Preprocessing:** Images and captions are processed using an AutoProcessor. Images are transformed into pixel embeddings, while captions are tokenized. A custom PyTorch dataset is used to enable efficient data loading and batching.
3. **Model Selection:** The BLIP Image Captioning Base model is selected, which combines a vision encoder and a text decoder to achieve multimodal understanding.
4. **Training:** The model is fine-tuned for multiple epochs using the AdamW optimizer. Cross-entropy loss is minimized through backpropagation by treating caption tokens as training labels.

#### Working:

1. The input image is first passed through the vision encoder to extract visual features.
2. These visual embeddings are then fed into the text decoder.
3. During training, the decoder learns to associate visual features with the correct sequence of caption tokens.
4. During inference, the trained model generates captions token by token using only the image as input.
5. The generated token IDs are decoded into a human-readable natural language caption.

#### Outcomes:

1. The model successfully generates relevant and football-specific image captions.
2. Caption accuracy improves significantly due to domain-specific fine-tuning.
3. The fine-tuned model is deployed on the Hugging Face Hub for easy reuse and sharing.
4. The project demonstrates the effectiveness of multimodal transformer architectures.

#### Conclusion:

This project demonstrates that fine-tuning transformer-based vision–language models can substantially enhance image captioning performance within a specific domain. The approach can be extended to other application areas such as medical imaging, wildlife monitoring, or surveillance systems.