Laboratory file

on

AGENTIC AI

SHARDA UNIVERSITY

Beyond Boundaries

School of Engineering and Technology

Department of Computer Science and Engineering

Subject code – **CSCR 3215**

SUBMITTED BY:

Name : Rachna Singh
System ID: 2023371832

SUBMITTED TO:

Mr. Ayush Singh

Sharda University

Greater Noida, Uttar Pradesh

Lab 01: Fine-Tuning

Finetune BLIP on an image captioning dataset

Objective:

The objective 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 understanding both visual and textual information.

Methodology:

1. Dataset Collection: Used a football image-caption dataset from Hugging Face containing images with corresponding textual descriptions.

2. Data Preprocessing: Images and captions are processed using AutoProcessor, where images are converted into pixel embeddings and captions are tokenized. A custom PyTorch dataset enables efficient batching.

3. Model Selection: BLIP Image Captioning Base model is employed, which integrates a vision encoder with a text decoder for multimodal understanding.

4. Training: The model is fine-tuned using the AdamW optimizer over multiple epochs. Cross-entropy loss is minimized through backpropagation using caption tokens as labels.

Working:

1. The input image is passed through the vision encoder, which extracts visual features.

2. These visual embeddings are provided to the text decoder.

3. During training, the decoder learns to map visual features to the correct caption tokens.

4. During inference, the trained model generates captions token-by-token based only on the image.

5. The generated token IDs are decoded into a readable natural-language sentence.

Outcomes:

1. The model successfully generates relevant football-specific captions.

2. Improved caption accuracy due to domain-specific fine-tuning.

3. The fine-tuned model is deployed on Hugging Face Hub for reuse.

4. Demonstrates effective use of multimodal transformers.

Conclusion:

This project proves that fine-tuning transformer-based vision-language models significantly improves image captioning performance in a specific domain. The system can be extended to other datasets such as medical, wildlife, or surveillance images.