Agentic AI Lab

(CSCR 3215)

BACHELOR OF TECHNOLOGY

In

Computer Science and Engineering

Submitted by:

Mayank Kumar

(2023542006) CSE-F G2

Submitted to: Mr. Ayush Singh

Assistant Professor

Department of Computer Science and Engineering

School of Computing Science & Engineering

Sharda University, Greater Noida

Jan-June 2026

Lab 1: Fine-Tuning BLIP on an Image Captioning Dataset

1. Introduction

Image captioning is a multimodal artificial intelligence task that combines computer vision and natural language processing to generate meaningful textual descriptions for images. Recent advancements in transformer-based architectures have significantly improved the performance of such systems. One prominent model in this domain is BLIP (Bootstrapped Language-Image Pretraining), which is designed to efficiently learn vision-language representations.

This experiment focuses on fine-tuning the BLIP model on an image captioning dataset to improve its ability to generate accurate and context-aware captions. Fine-tuning allows the pre-trained model to adapt to a specific dataset, thereby enhancing task-specific performance.

2. Objective

The objectives of this lab are:

• To understand the working of the BLIP model for image captioning.

• To fine-tune a pre-trained BLIP model on a custom image-caption dataset.

• To evaluate the effectiveness of fine-tuning in generating meaningful captions.

• To gain hands-on experience with multimodal deep learning using Hugging Face Transformers.

3. Dataset Description

The dataset used in this experiment consists of:

• A collection of images.

• Corresponding textual captions describing the content of each image.

Each data sample includes an image and one or more natural language descriptions. The dataset is preprocessed to ensure compatibility with the BLIP processor, including image resizing, normalization, and tokenization of text captions.

4. Methodology

The methodology followed in this experiment includes the following steps:

1. Model Selection

A pre-trained BLIP model is loaded from the Hugging Face Transformers library. This model has been previously trained on large-scale vision-language datasets.

2. Preprocessing

• Images are processed using the BLIP image processor.

• Text captions are tokenized using the BLIP tokenizer.

• Both image and text inputs are combined into a multimodal input format.

3. Fine-Tuning Process

• The model is trained using supervised learning.

• Loss is computed between predicted captions and ground-truth captions.

• Optimization is performed using gradient descent with appropriate learning rate and batch size.

4.Training Environment

The notebook is executed in Google Colab, leveraging GPU acceleration for efficient training.

5. Implementation Details

• Frameworks Used: PyTorch, Hugging Face Transformers

• Model: BLIP (Vision-Language Transformer)

• Execution Platform: Google Colab

• Optimization: Adam optimizer

• Loss Function: Cross-entropy loss for language modeling

The code is modular and follows standard deep learning training practices including checkpointing and evaluation.

6. Results and Observations

After fine-tuning, the BLIP model demonstrates:

• Improved caption relevance to image content.

• Better contextual understanding compared to the base pre-trained model.

• Ability to generalize reasonably well on unseen images.

Generated captions show enhanced semantic alignment with visual elements such as objects, actions, and scenes.

7. Conclusion

This experiment successfully demonstrates the fine-tuning of a BLIP model for image captioning. The results highlight the effectiveness of transfer learning in multimodal AI tasks. Fine-tuning pre-trained vision-language models significantly improves caption quality and contextual accuracy. This lab provides valuable insights into real-world applications of multimodal deep learning, such as assistive technologies, content generation, and image understanding systems.