Agentic AI Lab

(CSCR 3215)

B.TECH 3rd YEAR

SEMESTER: 6th

SESSION: 2025-2026

Submitted By:

Bhavishya Bhardwaj

(2023498559)

SECTION: H (G2)

Submitted To

Mr. Ayush
Kumar Singh
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SHARDA SCHOOL OF ENGINEERING & TECHNOLOGY

SHARDA UNIVERSITY, GREATER NOIDA

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

Aim

To fine-tune the **BLIP (Bootstrapping Language–Image Pre-training)** model on a custom image captioning dataset in order to improve the model's ability to generate accurate and context-aware captions for images.

Objective

- Understand multimodal learning using vision–language models
- Implement transfer learning using a pre-trained BLIP model
- Train the model on a custom dataset
- Evaluate image caption generation performance

Software & Tools Used

- **Programming Language:** Python
- **Platform:** Google Colab / Jupyter Notebook
- **Deep Learning Framework:** PyTorch
- **Libraries:**
    - transformers
    - datasets
    - torch
    - PIL
    - numpy
    - tqdm

Dataset Description

The dataset consists of:

- **Images**
- **Corresponding textual captions**

Each image–caption pair is used to train the model to associate visual features with natural language descriptions.

Model Used

BLIP (Bootstrapping Language–Image Pre-training)

BLIP is a vision-language model developed by Salesforce that supports:

- Image captioning
- Visual question answering
- Multimodal understanding

The pre-trained BLIP model is fine-tuned instead of training from scratch to reduce training time and improve accuracy.

Methodology / Algorithm

Step 1: Import Required Libraries

All necessary libraries for deep learning, dataset handling, and image processing are imported.

Step 2: Load Pre-trained BLIP Model and Processor

- BlipProcessor is used to preprocess both images and text.
- BlipForConditionalGeneration is loaded with pre-trained weights.

Step 3: Dataset Preprocessing

- Images are resized and normalized.
- Captions are tokenized.
- Image–caption pairs are converted into tensors suitable for training.

Step 4: DataLoader Creation

- The dataset is divided into batches.
- A PyTorch DataLoader is used for efficient training.

Step 5: Model Fine-Tuning

- Loss is calculated between predicted captions and actual captions.

- Backpropagation is applied to update model weights.
- Training is performed for multiple epochs.

Step 6: Model Evaluation

- The trained model generates captions for unseen images.
- Generated captions are compared with expected outputs to assess performance.

Code Explanation (High-Level)

- **Processor:** Converts raw images and text into model-compatible inputs
- **Model:** Generates captions using conditional language modeling
- **Optimizer:** Updates model parameters during training
- **Loss Function:** Measures caption generation error
- **Training Loop:** Repeats forward pass, loss calculation, and backpropagation

---

Results

- The fine-tuned BLIP model generates more relevant and descriptive captions compared to the base model.
- Training loss decreases over epochs, indicating effective learning.

Applications

- Automatic image captioning
- Assistive technologies for visually impaired users
- Content moderation and image indexing
- Multimedia search systems

Advantages

- Requires less data due to transfer learning
- Supports multimodal understanding
- High-quality caption generation

Limitations

- Computationally expensive
- Performance depends on dataset quality
- May generate biased captions if data is biased

Conclusion

In this experiment, a pre-trained BLIP model was successfully fine-tuned on an image captioning dataset. The model demonstrated improved caption generation performance, highlighting the effectiveness of transfer learning in multimodal deep learning tasks.

Future Scope

- Training on larger and more diverse datasets
- Fine-tuning for multilingual captioning
- Integration with real-time applications