

SHARDA UNIVERSITY

**DEPARTMENT OF COMPUTER SCIENCE &
ENGINEERING**

**SHARDA SCHOOL OF COMPUTING SCIENCES &
ENGINEERING**

January 2026

Fine-Tuning BLIP for Image Captioning

Enhancing Vision-Language Models for Custom Datasets

Presented by:

Abdullah Rayyan

Under the Supervision of:

Mr. Ayush Kumar Singh

Contents

1	1. Introduction	2
2	2. Problem Statement	2
3	3. Preliminary Literature Survey	2
4	4. Tools / Tech Stack	2
5	5. Data Sources	3
6	6. Aim (Broad Objectives)	3
7	7. Expected Outcomes	3
8	8. References	3

1. INTRODUCTION

The intersection of Computer Vision (CV) and Natural Language Processing (NLP) has led to the development of powerful Vision-Language models capable of understanding and generating content based on visual inputs. This project focuses on **BLIP (Bootstrapping Language-Image Pre-training)**, a state-of-the-art framework designed for unified vision-language understanding and generation.

Specifically, this project involves fine-tuning a pre-trained BLIP model on a custom image captioning dataset. By adjusting the model's weights, we aim to improve its ability to generate accurate, context-aware textual descriptions for images in a specific domain.

2. PROBLEM STATEMENT

While general-purpose multimodal models are powerful, they often lack the specificity required for domain-specific tasks.

- **Generic Descriptions:** Pre-trained models often provide safe, generic captions that miss crucial details relevant to specific applications (e.g., medical imaging, sports analysis, or e-commerce).
- **Domain Adaptation:** There is a need to adapt these large models to specialized datasets without training them from scratch, which is computationally expensive.
- **Performance Optimization:** Standard models may not perform optimally on datasets with unique vocabulary or visual styles.

3. PRELIMINARY LITERATURE SURVEY

- **BLIP Framework:** Proposed by Li et al., BLIP utilizes a multimodal mixture of encoder-decoder (MED) architecture. It effectively bootstraps from noisy web data by filtering and generating captions.
- **Transformers in Vision:** The shift from CNNs to Vision Transformers (ViT) has allowed for better alignment between image patches and text tokens.
- **Transfer Learning:** Fine-tuning pre-trained large models (like BERT or ViT) is a standard practice to achieve state-of-the-art results with limited labeled data.

4. TOOLS / TECH STACK

The project relies on the following technologies:

- **Programming Language:** Python 3
- **Deep Learning Framework:** PyTorch
- **Model Library:** Hugging Face Transformers ('transformers')
- **Data Handling:** Hugging Face Datasets ('datasets')
- **Model Architecture:** BLIP (Bootstrapping Language-Image Pre-training)
- **Hardware:** GPU acceleration (CUDA) is required for efficient training.

5. DATA SOURCES

The project utilizes image-caption pairs for training and validation.

- **Primary Dataset:** A specialized Image Captioning Dataset (typically loaded via the Hugging Face Hub, such as the 'ybelkada/football-dataset' or standard benchmarks like COCO/Flickr30k, depending on the specific notebook configuration).
- **Data Structure:** Each entry consists of an image file (pixel values) and a corresponding ground-truth text caption.

6. AIM (BROAD OBJECTIVES)

The broad objectives of this project are:

1. To implement a data pipeline that preprocesses images and tokenizes text for the BLIP architecture.
2. To fine-tune the 'Salesforce/blip-image-captioning-base' model on a new dataset.
3. To minimize the training loss effectively while preventing overfitting.
4. To demonstrate the model's ability to generate novel captions for unseen images after training.

7. EXPECTED OUTCOMES

- A fine-tuned BLIP model checkpoint saved and ready for inference.
- A comparative visualization showing "Before" vs. "After" captions, demonstrating the model's adaptation to the new dataset.
- A functional inference script that takes an arbitrary image as input and outputs a descriptive caption.
- Quantitative metrics (Training Loss) indicating successful convergence.

8. REFERENCES

- Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML.
- Hugging Face Documentation: https://huggingface.co/docs/transformers/model_doc/blip
- PyTorch Documentation: <https://pytorch.org/>