Image Captioning Using BLIP Model

1. Introduction

Image captioning is a multimodal task that combines **computer vision** and **natural language processing (NLP)** to generate meaningful textual descriptions for images. With the advancement of transformer-based architectures, vision-language models have significantly improved the quality of generated captions.

This project focuses on fine-tuning the **BLIP (Bootstrapping Language–Image Pre-training)** model on an image-captioning dataset to generate accurate and context-aware captions.

2. Objective

The main objectives of this project are:

- To understand the working of vision–language models
- To fine-tune a pre-trained BLIP model on a custom dataset
- To generate descriptive captions for input images
- To evaluate the performance of the fine-tuned model

3. Dataset Description

The dataset used in this project consists of:

- A collection of images
- Corresponding human-written captions for each image

Each image-caption pair helps the model learn visual concepts and their linguistic representations.

4. Model Overview: BLIP

BLIP (Bootstrapping Language–Image Pre-training) is a transformer-based multimodal model designed for vision-language tasks such as:

- Image Captioning
- Visual Question Answering (VQA)
- Image-Text Retrieval

It uses:

- A **Vision Encoder** to extract image features
- A **Text Encoder–Decoder** to generate captions

5. Methodology

5.1 Data Preprocessing

- Images are resized and normalized
- Captions are tokenized using a tokenizer
- Image-caption pairs are formatted for model input

5.2 Model Fine-Tuning

- A pre-trained BLIP model is loaded
- The model is fine-tuned using supervised learning
- Loss is calculated between predicted and ground-truth captions

5.3 Training Configuration

- Optimizer: AdamW
- Loss Function: Cross-Entropy Loss
- Training performed for multiple epochs

6. Working Principle

1. Input image is passed through the vision encoder
2. Visual features are extracted
3. Text decoder generates captions token by token
4. Model learns to align visual and textual representations

7. Results

- The fine-tuned model generates meaningful captions
- Caption quality improves with training epochs
- The model successfully generalizes to unseen images

8. Advantages

- Generates human-like captions
- Works well with limited fine-tuning data
- Uses state-of-the-art transformer architecture

9. Limitations

- Requires GPU for efficient training
- Performance depends on dataset quality
- Training time increases with dataset size

10. Applications

- Assistive technologies for visually impaired users
- Automated image tagging
- Content moderation
- Multimedia search engines

11. Conclusion

This project demonstrates the successful fine-tuning of a BLIP model for image captioning. The model effectively learns the relationship between visual features and natural language, producing accurate and context-aware captions.

12. Future Scope

- Training on larger and more diverse datasets
- Integration with real-time applications
- Multilingual caption generation
- Evaluation using BLEU, METEOR, and CIDEr scores