Agentic AI Lab

CSCR3215

B.Tech. (CSE)-VI Semester

Submitted to:
Mr Ayush Kumar

Submitted by:
Abhinav
(2023378721)

School of Engineering & Technology
Department of Computer Science & Engineering

Working of Fine-Tuning BLIP on an Image Captioning Dataset

1. Objective

The goal of this code is to fine-tune a pretrained BLIP (Bootstrapped Language Image Pretraining) model so that it can generate image captions specific to a custom dataset.

2. Libraries Used

PyTorch is used for training, Hugging Face Transformers provides the BLIP model, PIL handles images, and Pandas manages dataset annotations.

3. Dataset Loading

The dataset consists of image paths and their corresponding captions stored in a CSV file. Each row maps one image to one caption.

4. Custom Dataset Class

A PyTorch Dataset class loads images and captions. The BLIP processor converts them into tensors that the model can understand.

5. BLIP Processor and Model

The processor prepares both images and text, while the BLIP model generates captions conditioned on image features.

6. DataLoader

The DataLoader feeds data to the model in small batches, improving training efficiency and memory usage.

7. Training Setup

The AdamW optimizer updates model weights. The model is set to training mode to enable gradient updates.

8. Training Loop

For each batch, the model performs a forward pass, calculates loss, performs backpropagation, and updates weights using the optimizer.

9. Loss Function

The loss measures the difference between generated captions and ground-truth captions. Lower loss means better caption quality.

10. Saving the Model

After training, the fine-tuned model and processor are saved for future inference.

11. Inference

Given a new image, the model generates a caption which is decoded into human-readable text.

12. Conclusion

This process customizes a powerful pretrained BLIP model for domain-specific image captioning tasks.
