AGENTIC AI LAB

(CSCR3215)

BACHELOR OF TECHNOLOGY

In

Computer Science and Engineering

Department of Computer Science and Engineering

School of Computing Science & Engineering

Sharda University, Greater Noida

Jan-June 2026

Objective: Fine-tune the BLIP (Bootstrapped Language-Image Pretraining) model to generate captions for images from the Flickr30k dataset.

Steps and Observations:

1. **Environment Setup:**
    - Installed Hugging Face Transformers and Datasets library.
    - These are required for loading pre-trained BLIP and image-caption datasets.
2. **Dataset Loading:**
    - Dataset: Flickr30k (images + captions).
    - Loaded training split using load_dataset.
    - Sample output:
        - Image:
        - Caption: "A man riding a horse in the field."
3. **Preprocessing:**
    - Used AutoProcessor to convert images → tensors and captions → token IDs.
    - Created ImageCaptioningDataset class to feed data to the model.
4. **Model Loading:**
    - Pre-trained BLIP model (blip-image-captioning-base) loaded.
    - Model moved to GPU (if available) for faster training.
5. **DataLoader:**
    - Batch size = 4, shuffling enabled.
    - Enables efficient mini-batch training.
6. **Training:**
    - Optimizer: AdamW, learning rate = 5e-5.
    - Loss function: Cross-Entropy (built into BLIP).
    - 1 epoch training sample loss values:

Loss: 3.02

- Loss: 2.87
- Loss: 2.65
- Loss: 2.40
-

- Loss gradually decreases, indicating learning.

7. **Inference / Caption Generation:**
    - Evaluated model on sample images:

Input Image: [image of a man riding a horse]

- Generated Caption: "A person riding a horse in a grassy field."
-

- Captions are coherent and semantically match images.

Conclusion:

- Fine-tuned BLIP successfully generates captions for new images.
- Training reduces loss and improves caption quality.
- Approach can be extended for custom datasets or multiple epochs for better results.