Report Summary

Overview

This assignment focused on finetuning the BLIP (Base) image-captioning model on the Flickr30k dataset in order to generate more accurate and descriptive image captions. The workflow involved dataset preprocessing, configuring the BLIP model for multimodal training, executing multi-epoch finetuning, and evaluating model performance through both qualitative inspection and quantitative metrics.

Findings

Finetuning the BLIP model yielded several notable improvements:

- **Model Adaptation:** The finetuned model aligned much more closely with the Flickr30k captioning style, producing captions that were contextually richer and more precise.
- **Caption Quality Improvements:**
  - Better recognition of objects, scenes, and interactions.
  - Improved descriptions of human–object activities.
  - More detailed and dataset-specific captions compared to the baseline model.
- **Training Progress:** The training loss decreased steadily across all three epochs, indicating stable convergence and effective learning throughout the finetuning process.

Discussion of Evaluation Metrics

To assess the impact of finetuning, BLEU and ROUGE scores were computed for both the pretrained baseline and the finetuned BLIP model. Figures 1 and 2 summarize these results.

BLEU Score Comparison

The BLEU score comparison reveals consistent improvements across all evaluated BLEU variants (BLEU-1 through BLEU-4). The finetuned model achieved higher mean scores for every metric, indicating stronger alignment with human-generated captions. Improvements in higher-order BLEU scores—which emphasize longer, more structured n-gram matches—suggest that finetuning helped the model capture multi-word phrases and complex actions rather than relying on short, generic predictions.

ROUGE Score Comparison

A similar pattern is observed in the ROUGE metrics. The finetuned model outperformed the baseline in ROUGE-1, ROUGE-2, and ROUGE-L, demonstrating better recall of key descriptive phrases and improved reproduction of longer subsequences found in reference captions. Since ROUGE emphasizes recall, these gains suggest that the model learned to express more of the essential descriptive content present in each image's ground-truth caption.

Overall Interpretation

Together, the BLEU and ROUGE results strongly reinforce the qualitative improvements observed in example captions. The finetuned BLIP model not only generated captions that were more detailed and contextually appropriate but also achieved measurably better alignment with human descriptions. These findings show that domain-specific finetuning greatly enhances captioning performance compared to relying solely on the pretrained model.

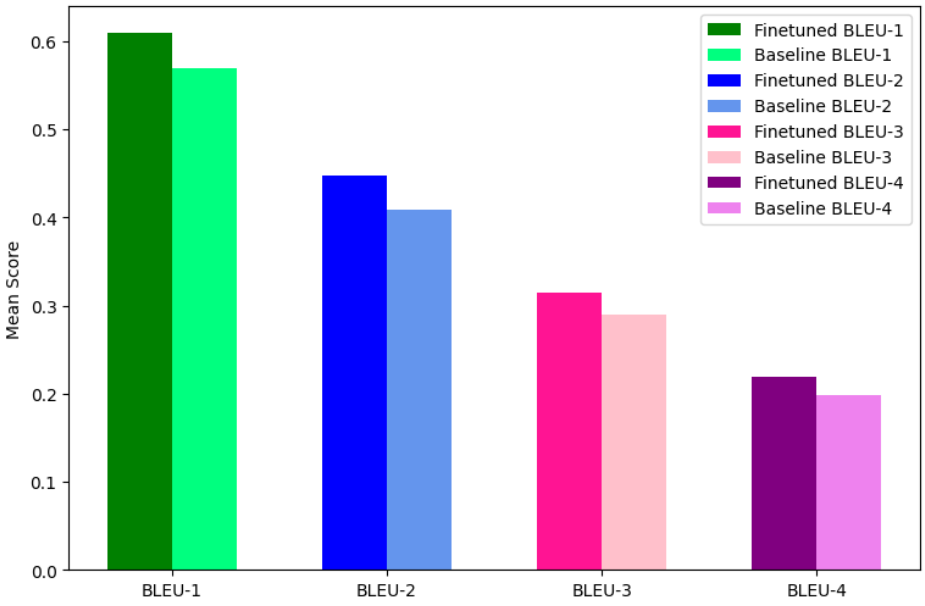

Figure 1: Comparison of BLEU scores between the baseline BLIP model and the finetuned model.

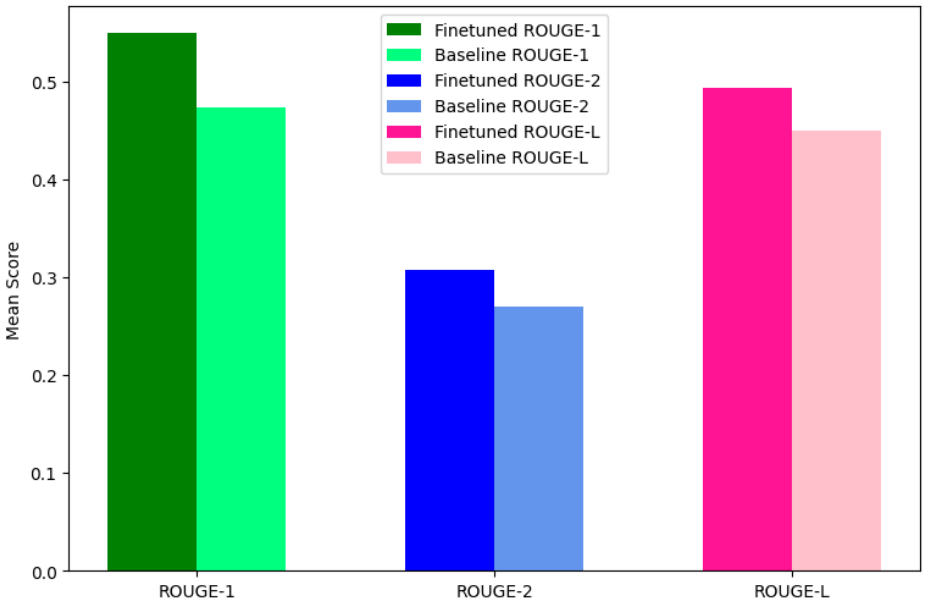

Figure 2: Comparison of ROUGE scores between the baseline BLIP model and the finetuned model.

Challenges Faced

Several challenges arose during this assignment:

- **Dataset Loading and Format Issues:** The Flickr30k dataset required custom handling due to outdated HuggingFace loading scripts, necessitating manual parsing and preprocessing.
- **Hardware Limitations:** Local hardware proved insufficient for finetuning such a large multimodal model. Training was therefore performed on Google Cloud Vertex AI Workbench using an NVIDIA L4 GPU. Even with this high-performance configuration, completing three epochs required approximately 1.5 hours.
- **Resource Management:** Proper batch sizing, mixed-precision training, and dataloader optimization were essential to prevent memory exhaustion and ensure smooth training.

Conclusions

Finetuning the BLIP model on the Flickr30k dataset led to substantial improvements in caption quality, descriptive accuracy, and alignment with human annotations. The use of Google Cloud Vertex AI with an NVIDIA L4 GPU was critical in enabling practical training times and overcoming computational constraints. Overall, this assignment provided valuable experience with multimodal model finetuning, cloud-based training workflows, dataset management, and performance evaluation using established NLP metrics.