Image-to-Text Generation: A Vision Language Model for Automated Image Captioning

Presented by

Saima Sanjida Shila
Master’s Student, Computer Science
Louisiana State University
sshila1@lsu.edu

Course: CSC7700-Foundational AI

Motivation:

- Manual captioning of large image datasets is time-consuming; AI-generated captions can make this process faster and more efficient.
- Accurate image captions help improve accessibility, content moderation, and human-AI interaction in visual applications.

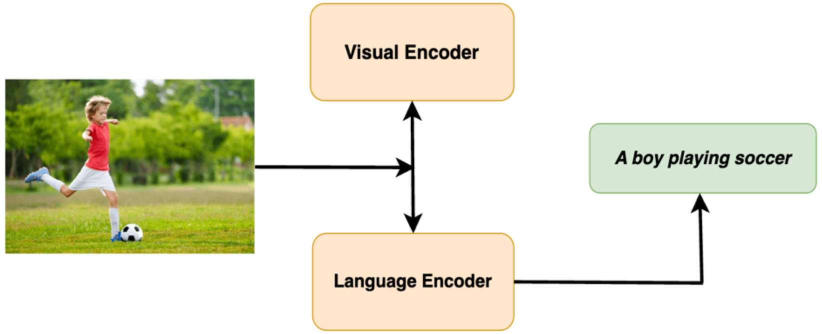
Objective:

- To fine-tune the BLIP model on the Flickr30k dataset for generating accurate and meaningful image captions.
- To develop a user-friendly web interface that allows users to upload images and receive real-time AI-generated captions.

BLIP Architecture:

- BLIP combines a visual encoder (like ViT or ResNet) with a language decoder (Transformer) to understand and describe images.
- It aligns visual features with text tokens to generate accurate, context-aware image captions.

BLIP Architecture



Dataset Used:

- Flickr30K which contains over 30,000 images, each paired with five human-written captions.
- The dataset was split into training (5,000), validation (1,000), and test (1,000) image-caption pairs.

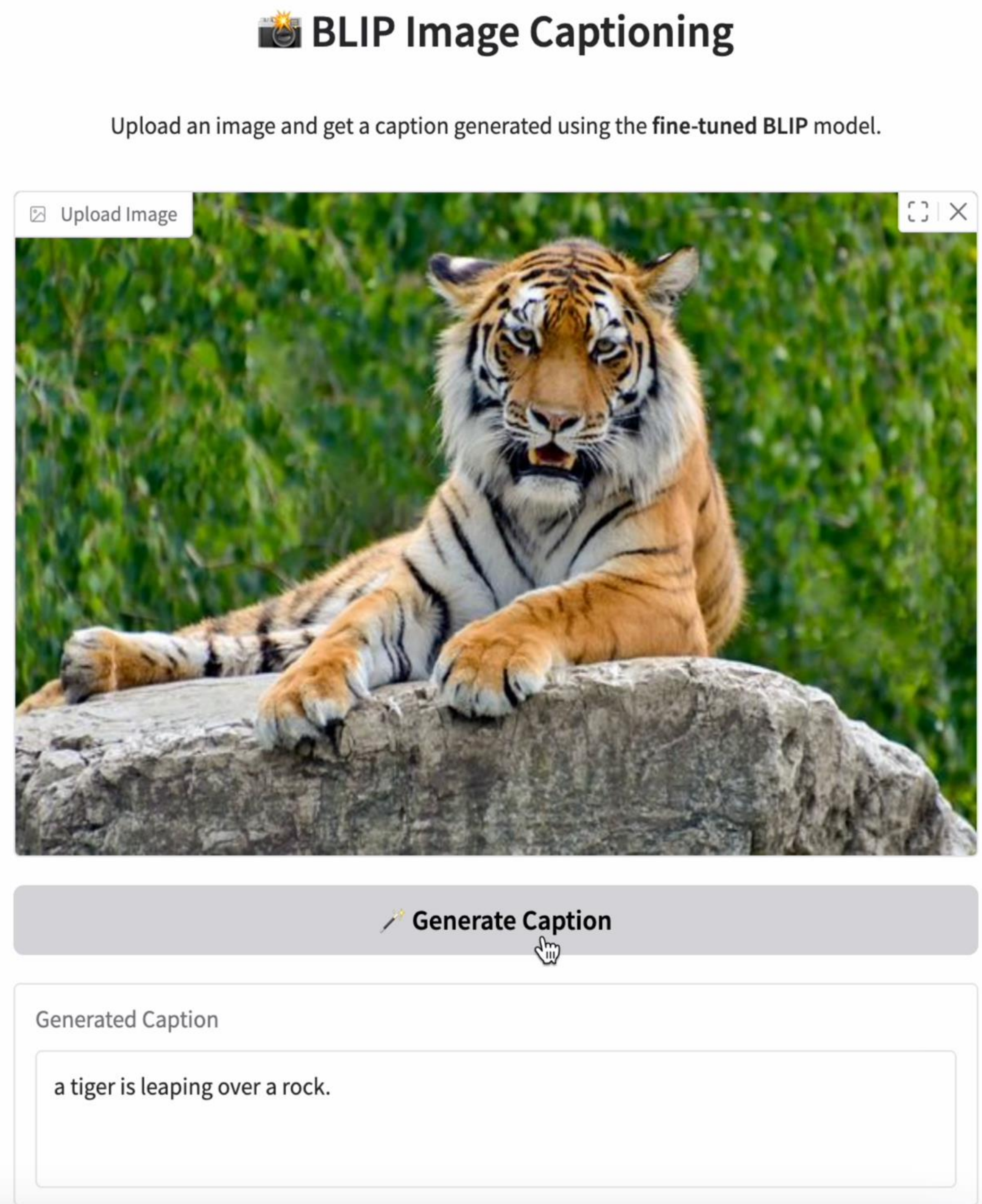
https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

Methodology:

- Preprocessed the Flickr30k dataset by converting images to RGB and tokenizing captions using the BLIP processor.
- Fine-tuned the pre-trained BLIP model on 5,000 training samples using the Hugging Face Trainer API.
- Trained for 5 epochs with a batch size of 2, using mixed precision (FP16).
- Model checkpoints saved every 500 steps monitoring training progress.
- Model performance was Evaluated the model on a 1,000-image test set using BLEU and ROUGE scores to measure caption quality.

User interface:

- A web-based interface is developed to allow users to upload images and receive real-time captions.
- The interface uses the fine-tuned BLIP model to generate and display captions instantly.



📸 BLIP Image Captioning

Upload an image and get a caption generated using the **fine-tuned BLIP** model.

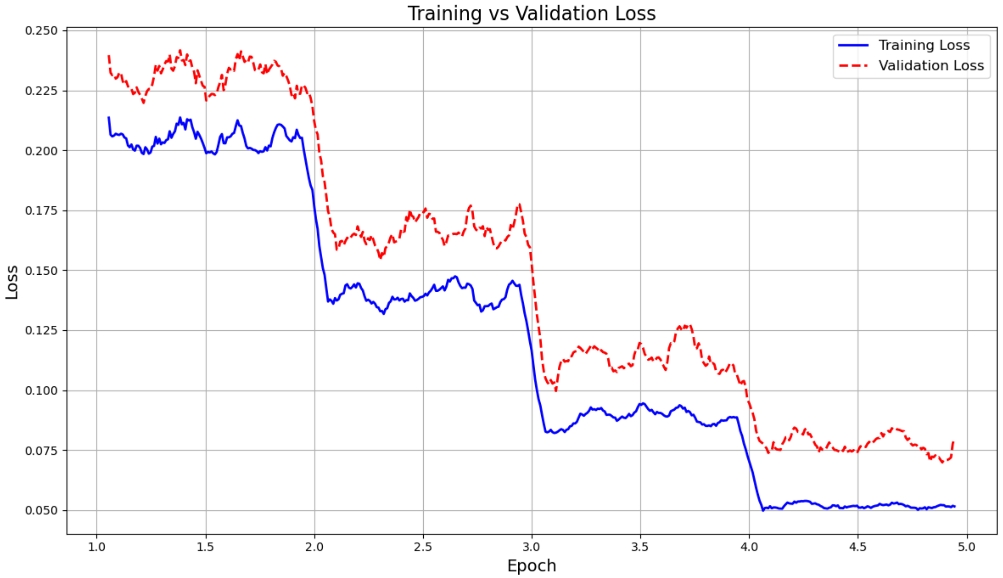
Generated Caption

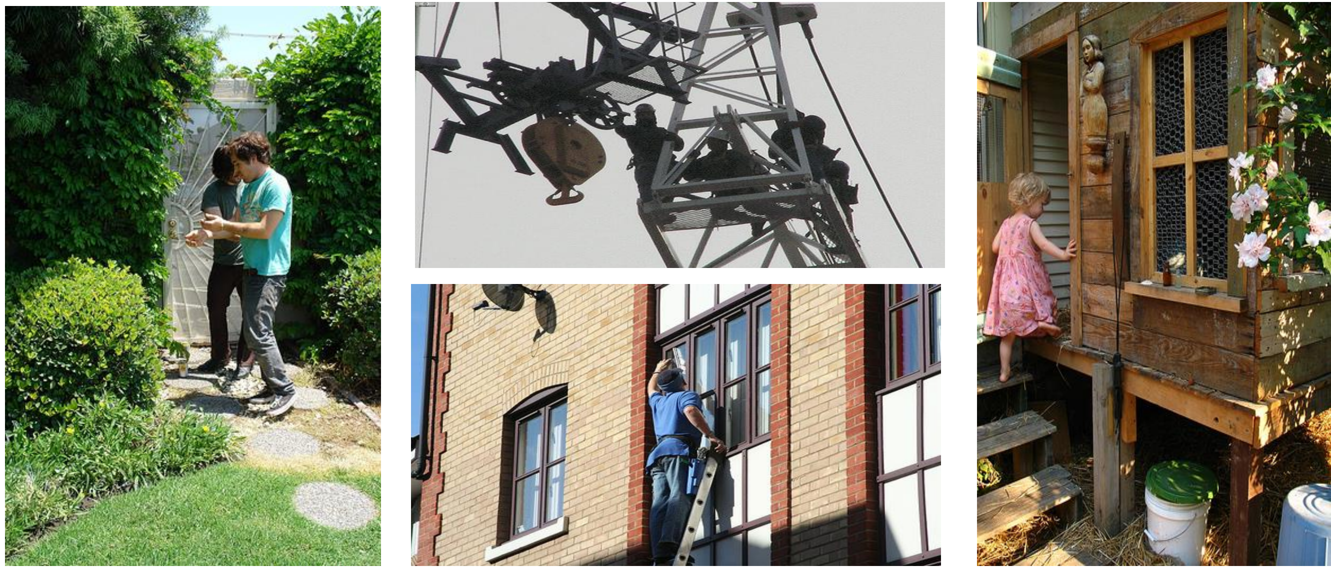
a tiger is leaping over a rock.

Results:

- Average BLEU score on test dataset: 0.0498
- Average ROUGE score on test dataset: 0.2566

Training/Validation Loss Curve:





Images are taken from Flickr30k dataset.

Actual Image Description	BLIP generated Caption
Two young guys with shaggy hair look at their hands while hanging out in the yard.	a man standing in the grass.
Two men working on a machine wearing hard hats.	a metal tower.
A little girl in a pink dress going into a wooden cabin.	a little girl in a pink dress.
man in blue shirt and jeans on ladder cleaning windows.	a man on a ladder.

Conclusion:

- This project successfully fine-tuned the BLIP model for image captioning using the Flickr30k dataset.
- The model achieved reasonable BLEU and ROUGE scores, showing its ability to generate meaningful captions.
- The real-time web interface demonstrates the practical potential of vision-language models in user-facing applications.

Future work:

- Expand into different dataset (MS COCO) to include more diverse and complex images for better generalization.
- Incorporate reinforcement learning with human feedback to improve caption quality.
- Use LLaVA (Llama + Vision Encoder) to train, test and evaluate and compare both model performance.

Thank You!