

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

**Abstract:** This project highlights the effectiveness of automated image captioning using Bootstrapping Language-Image Pretraining(BLIP), a state-of-the-art vision language model. The primary objective is to generate accurate and meaningful natural language descriptions of images by fine-tuning the pretrained BLIP model on widely used Flickr30k dataset. I have generated a user interface where users are allowed to upload an image, and instantly receive captions generated by the fine-tuned BLIP model. To evaluate the model performance, the model was evaluated using standard natural language generation metrics, achieving an average BLEU score of 0.0498 and a ROUGE score of 0.2566. These results indicate a decent level of precision and semantic similarity between the generated and reference captions. Overall, the project demonstrates both the feasibility and effectiveness of integrating advanced multimodal models into interactive tools that can enhance human-AI collaboration in image understanding tasks.

**Methodology:**

- 1. Dataset and preprocessing:** I have used Flickr30k dataset as the primary source for training, validation and testing. This widely used dataset contains over 30,000 images, each paired with five human-written captions. To prepare the dataset, captions were tokenized using tokenizer built into the BLIP framework. After preprocessing, images and captions were split into training, validation and testing datasets.

Training Dataset	5,000 image-caption pairs
Validation Dataset	1,000 image-caption pairs
Test Dataset	1,000 image-caption pairs

**Preprocessing steps:**

- Images are converted to RGB format using PIL library.
  - Each image-caption pair is tokenized using Blip Processor with a maximum length of 128 tokens.
  - Labels are assigned by cloning the input IDs which supports supervised learning.
- 2. Training and Evaluation:** The pre-trained BLIP model was fine-tuned on the training dataset of Flickr30k to better adapt it for the image captioning task. The main focus during training was on reducing captioning loss and improving the model’s ability to generalize to new, unseen image-caption pairs. The training ran over 5 epochs with real-time evaluations. A small batch size was selected to accommodate the limitations of available GPU memory. Over time, the training loss decreased over

epochs, showing that the model was learning effectively. The training was conducted using the Trainer API from Hugging Face Transformers with the following parameters.

- Batch Size: 2 (train and validation)
- Epochs: 5
- Checkpoint Save Steps: 500
- Save Limit: 2 most recent checkpoints
- Mixed Precision (FP16): Enabled if CUDA is available

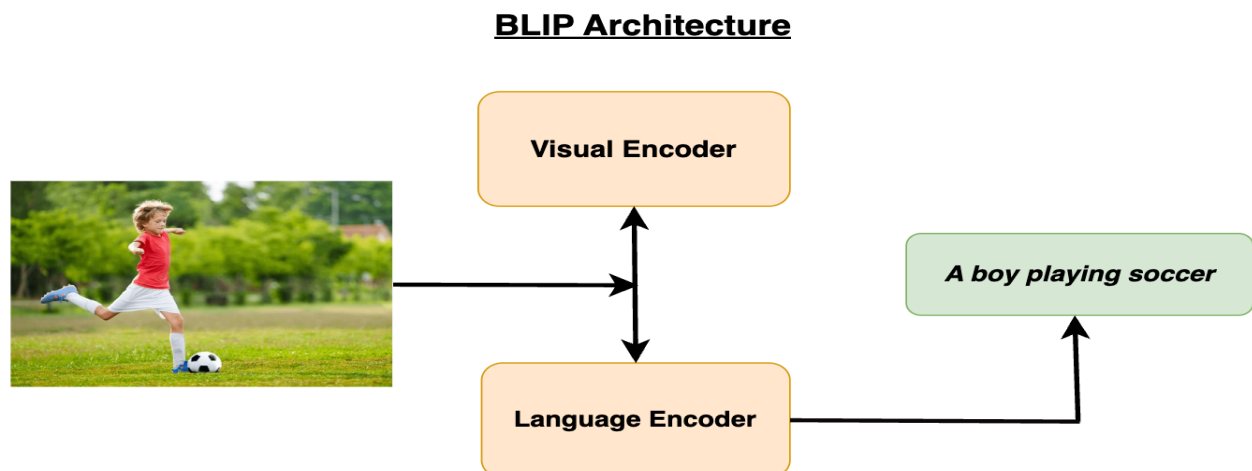
After training, the fine-tuned model was tested on a test dataset of 1,000 data images using standard evaluation metrics for image captioning.

Evaluation metrics:

BLEU : Measures how much the generated captions overlap with the reference captions at the n-gram level (short sequences of words).

ROUGE : Evaluates the similarity of word sequences and the longest common subsequences between generated and reference captions.

**3. Model Architecture:** The project used the BLIP (Bootstrapping Language-Image Pretraining) architecture, a vision-language model built to align visual and textual embeddings for image captioning tasks. BLIP combines a visual encoder (typically ViT or ResNet) with a transformer-based language decoder. This configuration allows the model to effectively learn cross-modal representations and generate accurate, context-aware image captions. Fine-tuning on the Flickr30k dataset improved the model's performance without requiring architectural modifications.




#### 4. User Interface


To enable real-time user interaction, I developed a simple Web Interface. The user interface allows users to upload an image, which is then processed by the fine-tuned BLIP model running in the backend. The model generates a caption that is immediately displayed to the user. The interface was designed to be intuitive and accessible, offering a seamless experience that showcases how multimodal AI systems can be effectively integrated into real-world applications for describing visual content.

### BLIP Image Captioning

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

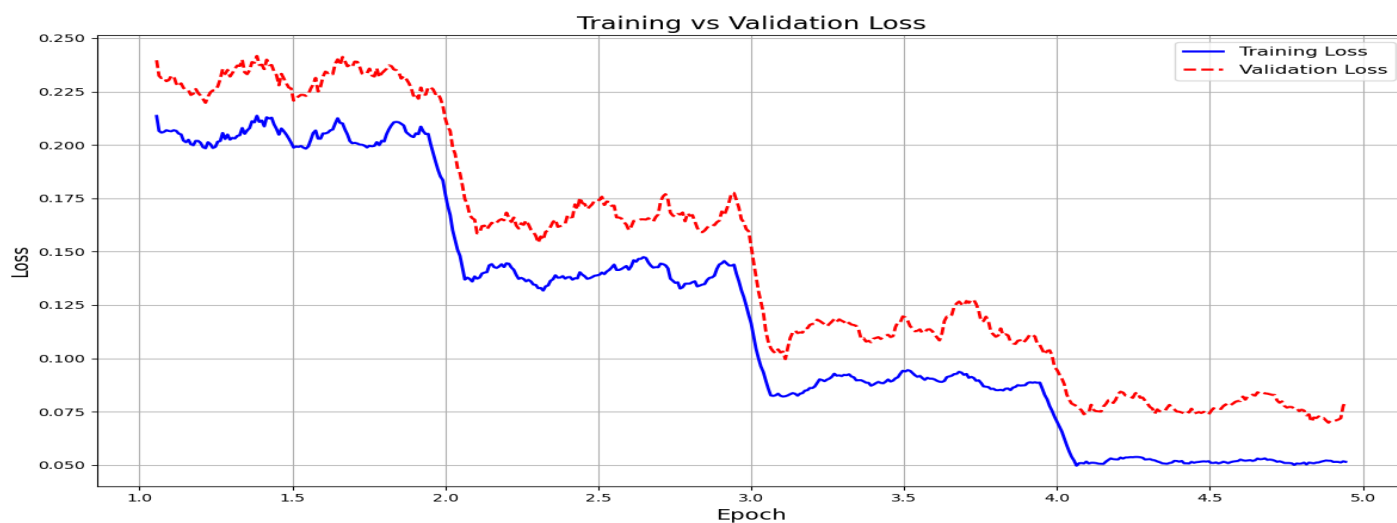
Upload Image



 **Generate Caption**

Generated Caption

a tiger is leaping over a rock.

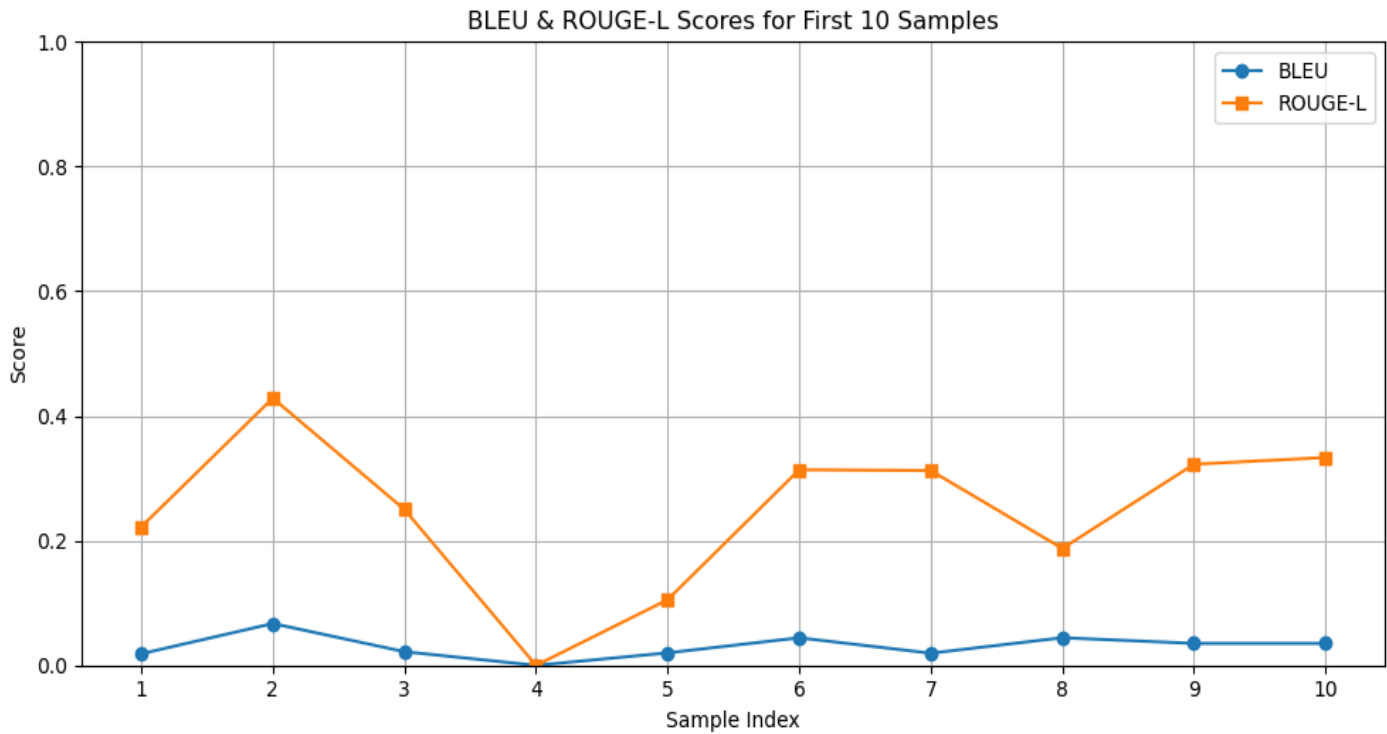
**Results:**Avg BLEU SCORE ON TEST DATASET: 0.0498AVG ROUGE SCORE ON TEST DATASET: 0.2566Training/Validation Loss curve:Comparison of Actual Human Written Caption & BLIP Generated Caption:

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.



Images are taken from Flickr30k dataset.

BLEU & ROUGE score for first 10 Samples:



Quad Chart:

<p><b><u>Challenge:</u></b></p> <ul style="list-style-type: none"> <li>Creating manual image captioning can be slow and inconsistent.</li> <li>Accessibility, content moderation, and robotic navigation systems need reliable, real-time image captioning.</li> </ul>	<p><b><u>Proposed Technical Solution:</u></b></p> <ul style="list-style-type: none"> <li>Fine-tune BLIP (Bootstrapped Language-Image Pretraining) model on Flickr30k dataset.</li> <li>Build a web-based UI with deployed fine-tuned BLIP model for real-time image captioning.</li> </ul>
<p><b><u>Impact:</u></b></p> <ul style="list-style-type: none"> <li>Accelerate image tagging for large image databases and social media.</li> <li>Enable safer and smarter AI decision-making in robotic navigation.</li> <li>Advance multi-modal AI development combining vision and language.</li> </ul>	<p><b><u>Benefits:</u></b></p> <ul style="list-style-type: none"> <li>A fine-tuned BLIP model achieving BLEU = 0.0498, ROUGE = 0.2566.</li> <li>A live user interface allowing image upload and instant captioning.</li> </ul>

**Code Repository:** [GitHub Repo Link](#)

**Discussion & Conclusion:** This project demonstrates using the BLIP (Bootstrapping Language-Image Pretraining) model for automated image captioning is both feasible and effective. While the BLEU score of 0.0498 highlights the challenges of strict n-gram evaluation in natural language tasks, the average ROUGE score of 0.2566 shows the model's ability to generate semantically meaningful captions. Adding a real-time user interface further boosts the project's practical relevance, illustrating how vision-language models can be integrated into interactive tools for everyday users.

Future work could include expanding the training dataset, using reinforcement learning with human feedback, or applying ensemble approaches to boost both accuracy and interpretability. Overall, this project makes a solid contribution to the evolving field of vision-language understanding and its application in smart, user-facing systems.