Marathwada MitraMandal's

COLLEGE OF ENGINEERING, PUNE

Accredited with 'A++' Grade by NAAC, Recipient of " Best College Award 2019" by SPPU

DEPARTMENT OF COMPUTER ENGINEERING

Mini Project Report

Sub: LP-VI (NLP)

Title: Image-to-Text Generation and Translation Tool

Problem Statement: The objective of this project is to develop an "Image-to-Text Generation and Translation Tool" using transformers. The tool aims to take an uploaded image as input and generate descriptive text captions for the content of the image in English. Additionally, it provides the functionality to translate these captions into different languages such as Hindi and Tamil, enhancing accessibility and understanding for a wider audience. The key components of the project include:

- Image Caption Generation: Utilizing state-of-the-art vision models to generate descriptive captions for uploaded images.
- Translation: Employing multilingual translation models to translate generated captions into various languages, enhancing the tool's accessibility and usability.
- User Interface: Implementing a user-friendly interface for seamless interaction and user experience.

Group Members:

BC201- Mayank Baber
BC211- Mayuresh Chougule
BC219 - Rachna Ganjoo

Software/Hardware Requirements:

- Software Requirements: Supporting OS(Windows/ MacOS/ Linux)
- Dependencies: Python, PyTorch, Transformers, Streamlit, and PIL
- Hardware Requirements: GPU(optional) for CUDA, CPU(8-core), RAM(12GB), 5GB Disk space

○

Detailed idea:

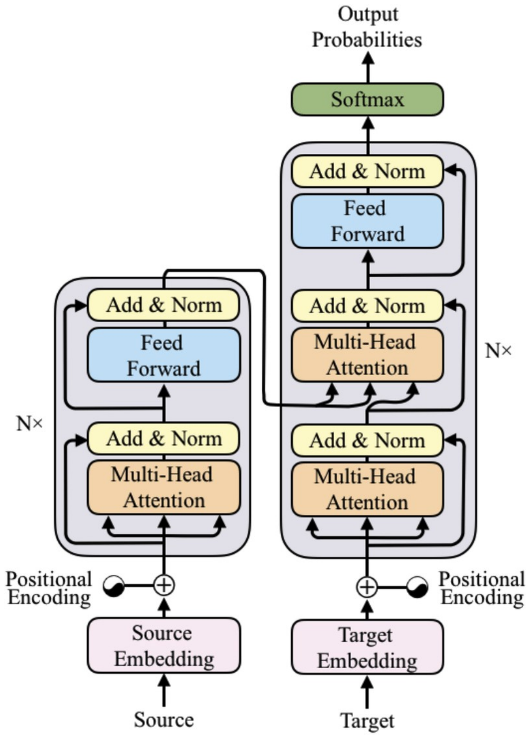
We have used the following models for image-to-text generation:

1) Visual Encoder-Decoder Model
2) Blip Model
3) Git-base COCO Model
4) mBART Model

1) **ViT Model**: The Vision Encoder Decoder Model can be used to initialize an image-to-text model with any pre-trained Transformer-based vision model as the encoder (e.g. ViT, BEiT, DeiT, Swin) and any pre-trained language model as the decoder (e.g. RoBERTa, GPT2, BERT, DistilBERT).

- Vision encoder-decoder models consist of two main components—an encoder that processes the input image and extracts meaningful features, and a decoder that generates textual descriptions based on these features.

- The decoder often utilizes Transformer-based architectures, such as BERT or GPT, to generate textual descriptions. Transformers are effective for generating text as they can capture long-range dependencies in the data.

- 

- Pre-trained encoder-decoder models, such as those trained on the COCO dataset, can be fine-tuned on specific image-to-text tasks to improve their performance on new datasets or domains.

2) **Blip Model**: BLIP models are trained on a large corpus of text and images. The text can come from sources like books, articles, and websites, while the images can be from datasets like COCO or ImageNet.

- BLIP models consist of both a language model (typically based on transformers) and an image encoder (such as a convolutional neural network). The model is trained to encode both text and image inputs and generate text outputs.

- After pre-training, the BLIP model is fine-tuned on a specific task, such as image captioning. Fine-tuning adapts the model to perform well on the target task using a smaller, task-specific dataset.

- The image encoder in the BLIP model converts images into embeddings, which are high-dimensional numerical representations. These embeddings capture the visual content of the image and are used as input to the model.

- During inference, the BLIP model takes an image as input, converts it into embeddings, and generates a textual description of the image.

3) **Git-base COCO Model**: The COCO dataset is commonly used for training image captioning models. It contains a large collection of images, each paired with multiple human-annotated captions describing the contents of the image.

    - The base-sized model refers to a pre-trained neural network architecture that serves as the foundation for the image-to-text conversion task. These models are typically based on transformer architectures like BERT or similar models adapted for image captioning tasks.

    - The base-sized model is fine-tuned on the COCO dataset to adapt it specifically for generating captions for images. Fine-tuning involves updating the weights of the pre-trained model using the COCO dataset to improve its performance on the image captioning task.

    - Once the model is trained and fine-tuned on the COCO dataset, it can be integrated into an NLP pipeline for image captioning. The model takes an image as input, processes it using the vision encoder, and generates a textual description of the image using the decoder.

4) **mBART Model**: This model is a fine-tuned checkpoint of mBART-large-50. It is fine-tuned for multilingual machine translation. It was introduced in Multilingual Translation with Extensible Multilingual Pretraining and Finetuning paper.

    - mBART is based on the BART (Bidirectional and Auto-Regressive Transformers) architecture, which combines the advantages of autoencoding and autoregressive models. It uses a masked language modelling objective for pretraining.

    - mBART can handle multiple languages. It can be fine-tuned on a wide range of languages, making it suitable for multilingual applications.

    - mBART is effective in understanding and generating text in languages it has been trained in, making it useful for tasks where the input language may vary.

    - mBART is pre-trained on a large corpus of text from various languages, which helps it capture language-specific nuances and improve performance on diverse language tasks.

    - The performance of mBART for image-to-text conversion can be improved through hyperparameter tuning and careful evaluation on a validation set.

Execution and results:

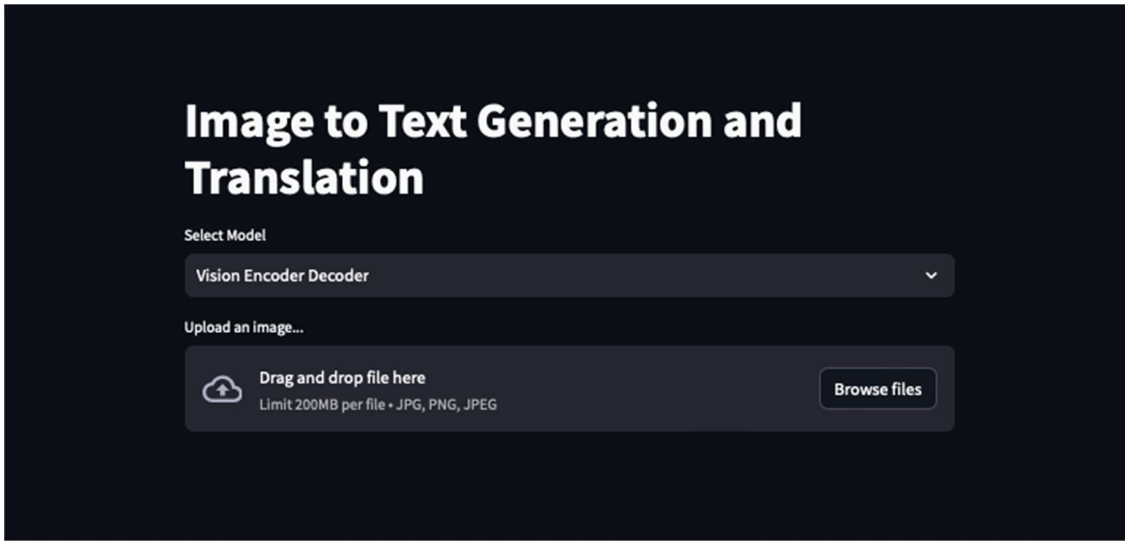


Image to Text Generation and Translation

Select Model: Vision Encoder Decoder

Upload an image...

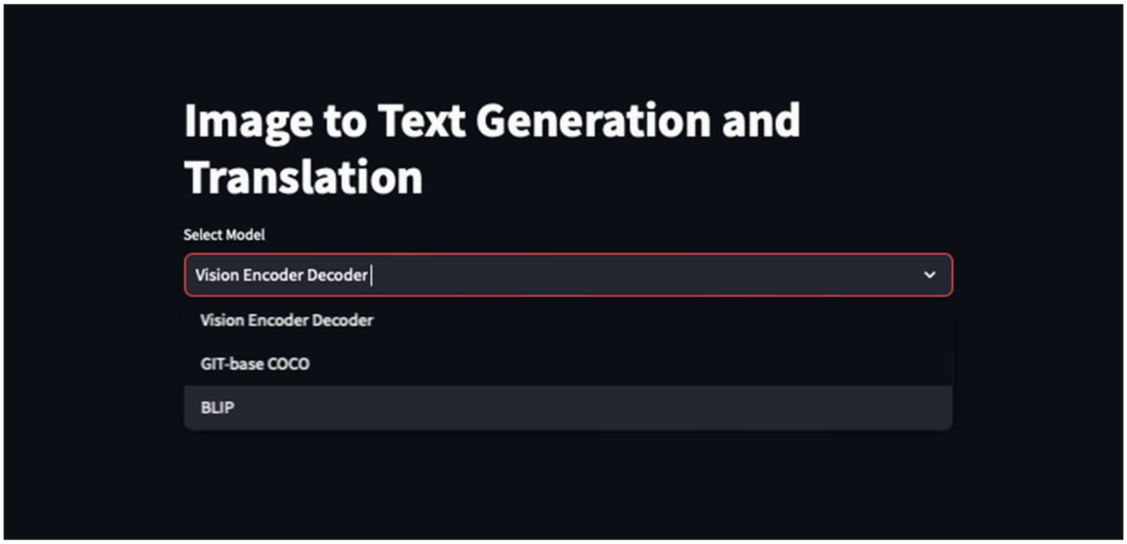


Image to Text Generation and Translation

Select Model: Vision Encoder Decoder

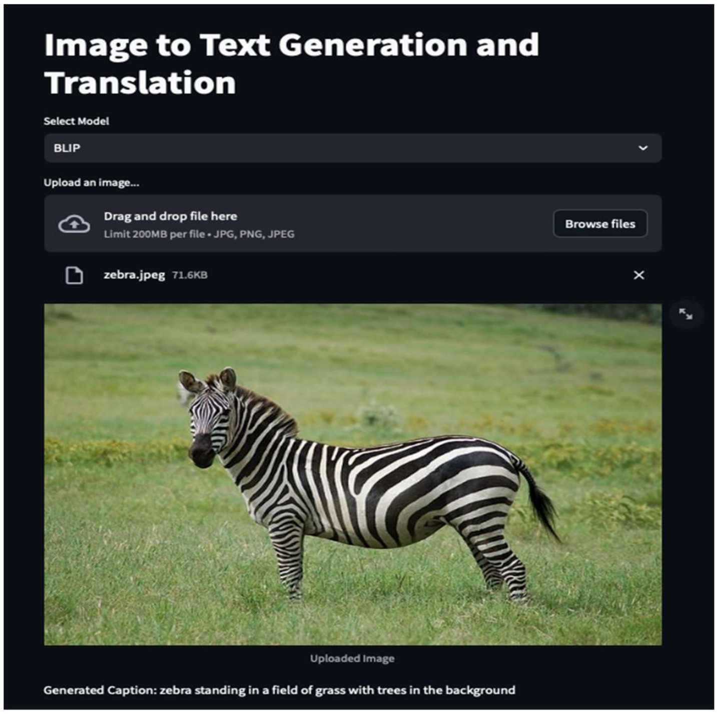


Image to Text Generation and Translation

Select Model

BLIP

Upload an image...

Generated Caption: zebra standing in a field of grass with trees in the background

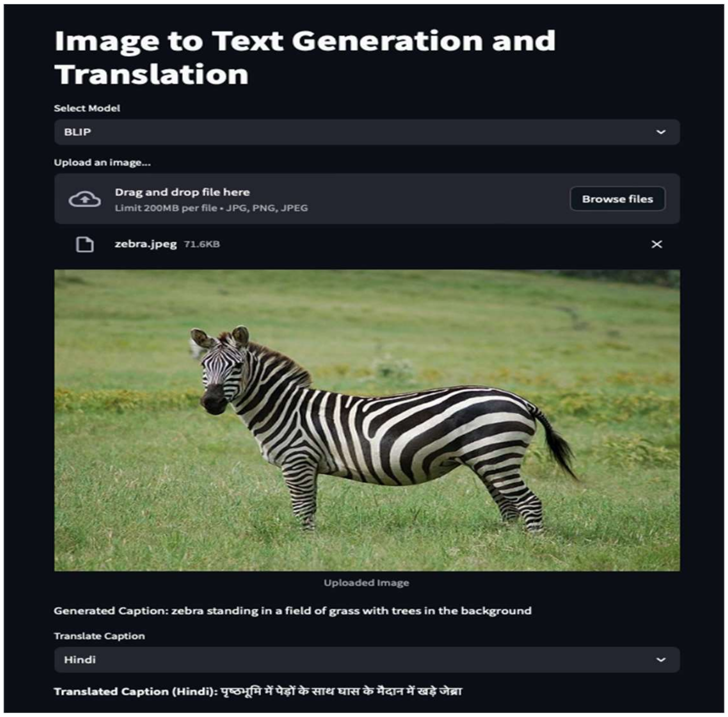


Image to Text Generation and Translation

Select Model

BLIP

Upload an image...

Generated Caption: zebra standing in a field of grass with trees in the background

Translate Caption

Hindi

Translated Caption (Hindi): पृष्ठभूमि में पेड़ों के साथ घास के मैदान में खड़े जेब्रा

Conclusion

This project showcases a simple yet effective way to generate captions from images using various pre-trained models. Leveraging the power of Vision Encoder-Decoder, BLIP, and GIT-base COCO models, it demonstrates the potential of AI in understanding visual content and translating it into textual descriptions. Additionally, it offers a translation feature to convert captions into Hindi or Tamil, highlighting the versatility of multilingual capabilities. While not state-of-the-art, this project serves as an accessible introduction to AI-driven image captioning and translation, making complex technologies more approachable for users.