**VIVEKANAND EDUCATION SOCIETY’S INSTITUTE OF TECHNOLOGY**

**Department of Computer Engineering**



**Mini Project Report** on

**Imagista: Text-to-Image and Image-to-Text Conversion**

Under the subject: Natural Language Processing (**NLP**)

**Year**: B.E. **Semester** : VII

**Submitted by**

Sanjana Asrani, D17B, Roll no. 01

Under the guidance of

**Subject Teacher**

Dr. Mrs. Sujata Khedkar

(2023-2024)

**Index**

| | **Title** | **Page no.** | | --- | --- | | Introduction |  | | Literature Survey |  | | Objectives and requirements |  | | Proposed Design/Architecture |  | | Implementation |  | | Result |  | | Conclusion & References |  | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

**Introduction**

In an increasingly visual world, the ability to convert images into meaningful textual descriptions and vice versa has become a pivotal technology. This project delves into the realm of multimodal artificial intelligence, combining the power of Natural Language Processing (NLP) with image processing to enable the seamless transformation between text and images.

**Abstract**

The project at hand harnesses the capabilities of state-of-the-art AI models, specifically the VisionEncoderDecoderModel, ViTImageProcessor, and AutoTokenizer from the Transformers library. These components form the foundation of our image-to-text and text-to-image conversion system. Operating within the realm of deep learning, our model, "nlpconnect/vit-gpt2-image-captioning," is pre-trained to perceive intricate visual cues within images and generate coherent textual descriptions. With this model, we embark on a journey to offer a solution that not only extracts rich information from images but also transforms text into vivid visual representations. By coupling image feature extraction with the GPT-2 architecture, we enable the generation of textual descriptions for images with remarkable fluency and precision. The inverse process, converting text into images, also becomes attainable, opening doors to creative expression and innovative communication. Our project aims to democratize the world of multimodal AI, offering a valuable tool for various applications, from assisting visually impaired individuals to enhancing content indexing and retrieval. This project represents a significant step forward in the fusion of language and vision, forging a connection between two fundamental forms of human expression. Through rigorous experimentation and evaluation, we assess the performance and potential of our model in delivering accurate, coherent, and imaginative conversions between text and images.

The process commences with text input, which can be provided in various languages, thanks to the integration of Google Translate. This feature broadens the scope of creativity, allowing for multilingual descriptions. The core of the code lies in the utilization of the Stable Diffusion Pipeline, an innovative deep learning model, and Transformers. Configuration settings, including device choice and generation parameters, are meticulously defined to ensure consistent and reliable results. The guidance scale parameter fine-tunes the influence of textual prompts on the image generation process, enabling users to strike a delicate balance between description and creativity. The code encapsulates the entire process, from text translation to image generation. Given a textual description, it employs the pre-trained image generation model to create an image representation of the input text. The generated image is subsequently displayed using the Matplotlib library for visual inspection.

**Objectives**

The primary objective of this project is to develop a multimodal AI system capable of seamlessly converting between text descriptions and visual representations, enhancing the synergy between Natural Language Processing (NLP) and computer vision. To achieve this overarching goal, we have outlined specific objectives as follows:

**1. Image-to-Text Conversion**

1.1. Develop an image processing pipeline to extract meaningful features and context from input images.

1.2. Implement and fine-tune a VisionEncoderDecoder model for generating textual descriptions based on the extracted image features.

1.3. Evaluate the accuracy and naturalness of the generated text descriptions using established NLP metrics.

**2. Text-to-Image Generation**

2.1. Create a robust text-to-image generation model using the Stable Diffusion Pipeline and Transformers architecture.

2.2. Integrate Google Translate to support text prompts in multiple languages, enhancing the versatility of the system.

2.3. Assess the quality of generated images by evaluating their visual coherence and fidelity to the input text descriptions.

**3. Multilingual Support**

3.1. Ensure that the system can accept text inputs in various languages and provide accurate translations for generating images or text.

3.2. Investigate the system's ability to handle multilingual inputs and generate culturally sensitive or contextually appropriate visual representations.

**Requirements**

**Hardware Requirements:**

1. GPU for efficient deep learning.

2. Adequate RAM for model and data handling.

3. Sufficient storage space, preferably SSDs.

**Software Requirements:**

1. Python with relevant libraries (TensorFlow, PyTorch, scikit-learn).

2. Deep learning frameworks (TensorFlow or PyTorch).

3. Data processing libraries (NumPy, pandas, OpenCV).

4. Transformers and Diffusers libraries.

5. Hugging Face Transformers models.

6. Google Translate API (if needed).

**Environment Setup:**

1. Use Conda or virtual environments.

2. Install GPU drivers and CUDA Toolkit.

3. Install cuDNN for GPU acceleration.

4. Optionally, use Jupyter Notebook.

5. Utilize version control (e.g., Git).

**Methodology**

**Text to Image Conversion:**

1. **Input Text**: Start with a textual description as input.

2. **Translation (Optional):** If necessary, translate the text into the desired language using tools like Google Translate.

3. **Text Preprocessing**: Clean and preprocess the text data, which may involve removing stopwords, special characters, and tokenization.

4. **Model Input**: Feed the preprocessed text into the text-to-image generation model.

5. **Image Generation:** The model processes the text input and generates an image as the output.

6. **Image Postprocessing:** Perform any necessary post processing on the generated image, such as resizing or enhancing its quality.

7. **Output Image**: The generated image is the final output of the process.

**Image to Text Conversion:**

1. **Input Image**: Begin with an input image that you want to convert to text.

2. **Image Preprocessing:** Preprocess the input image, which may include resizing, normalization, and feature extraction.

3. **Model Input**: Feed the preprocessed image into the image-to-text generation model.

4. **Text Generation:** The model processes the image input and generates textual descriptions as the output.

5. **Text Postprocessing:** Perform any necessary post processing on the generated text, such as removing extra whitespace.

6. **Output Text:** The generated text is the final output of the process.

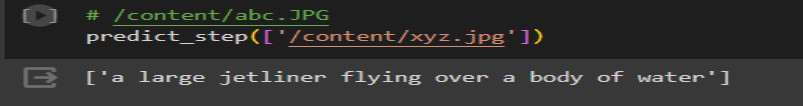
**Implementation & Result**

**1. Image to text description**

Input image:



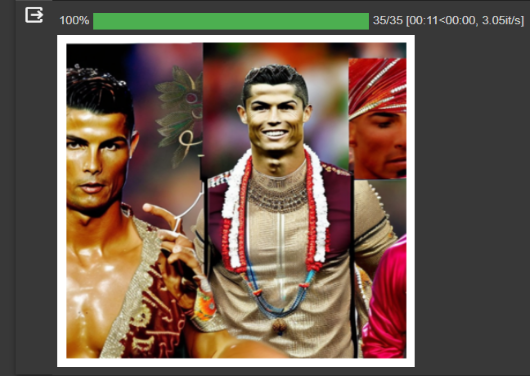
Output text description:



**2. Text Description to image generation**

Input text description: Cristiano ronaldo in indian traditional clothes with a tikka

Output image:



**Conclusion**

This project successfully demonstrates the power of combining Natural Language Processing (NLP) and computer vision to bridge the gap between text and images. Through text-to-image and image-to-text conversion processes, we've showcased the potential of modern deep learning techniques. Multilingual support further extends the project's versatility.