Vision Based Large Language Model for Visual Knowledge

Abstract

Our project introduces a novel method for translating GIFs and images using Moondream LLM and Vision Transformers. By combining Moondream's multimodal translation capabilities with Vision Transformers, we achieve accurate and coherent translations from images to natural language descriptions. Our algorithm processes GIFs frame by frame, ensuring temporal consistency. The aim of this project is to develop a free/open source image and gifs translator. Advantages of the model include semantic coherence, temporal consistency in GIF translation, and multimodal capabilities through Moondream and Vision Transformers.

Methodology

Large Language Models (LLMs):

Large language models (LLM) are very large deep learning models that are pre-trained on vast amounts of data.The foundation of our methodology lies in the utilisation of Large Language Models (LLMs) are capable of understanding complex linguistic patterns and generating coherent text. In our project, we leverage the capabilities of LLMs to facilitate the translation of visual content into natural language descriptions.

Moondream

Moondream is a Large Language Model for multimodal machine translation plays a pivotal role in our methodology. By integrating Moondream, enabling the model to comprehend and generate textual descriptions of visual content accurately

Vision Transformers

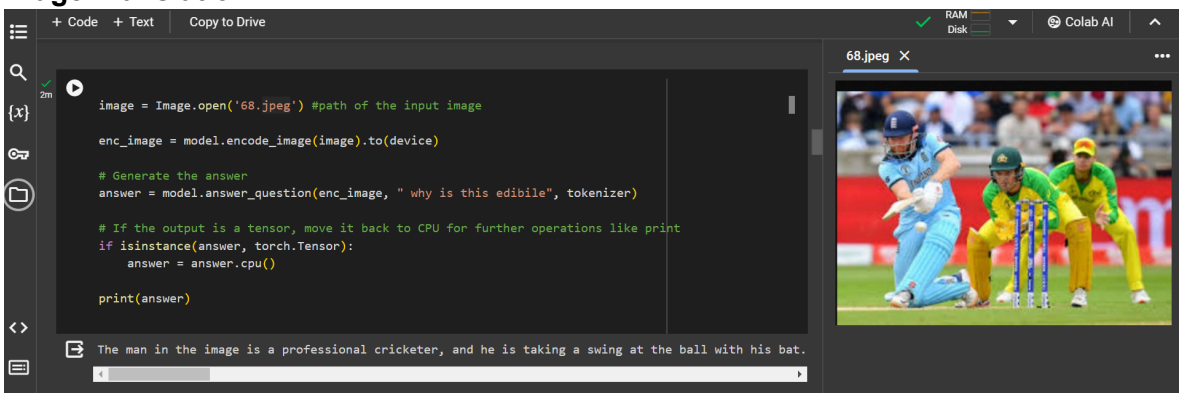
The ViT is a visual model based on the architecture of a transformer originally designed for text-based tasks. The ViT model represents an input image as a series of image patches, like the series of word embeddings used when using transformers to text, and directly predicts class labels for the image.

Frame-by-Frame GIF Translation:

To translate GIFs effectively, we adopt a frame-by-frame approach. Each frame of the GIF is treated as an individual image input to the algorithm. By processing GIFs frame by frame, we ensure temporal consistency and preserve the narrative coherence of the translated content. This methodology allows for accurate translation of dynamic visual sequences while maintaining semantic fidelity.By integrating these components into our methodology, we establish a robust framework for translating GIFs using Moondream and Vision Transformers. This approach enables us to achieve accurate and coherent translations from visual content to natural language descriptions, advancing the frontier of multimodal machine translation.

Results

Image Translation



Gifs Translation