

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

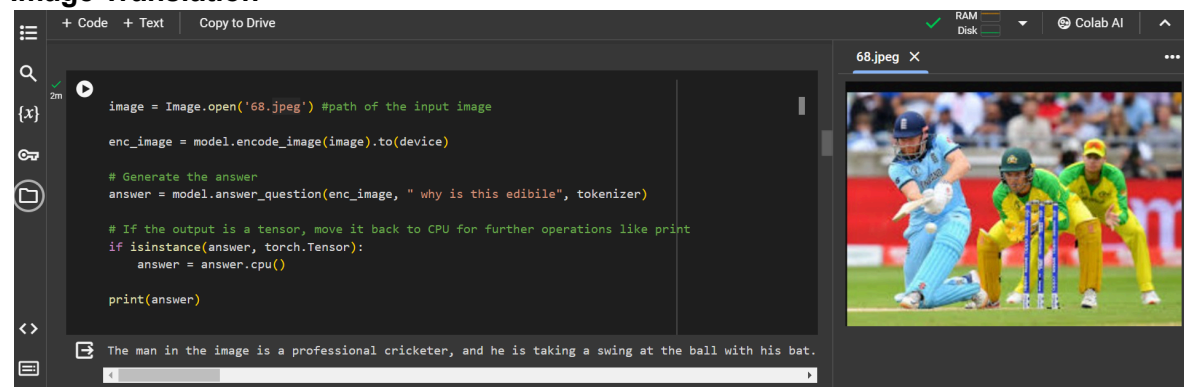
The ViT is a visual model based on the architecture of a transformer originally designed for text-based tasks. The ViT model represents an input image as a series of image patches, like the series of word embeddings used when using transformers to text, and directly predicts class labels for the image.

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



The screenshot displays a Google Colab notebook interface. The left pane shows Python code that loads an image named '68.jpeg', encodes it, and uses a model to generate an answer to the question 'why is this edible?'. The right pane shows the image of a cricket player in a blue jersey swinging a bat, and the generated text output below it.

```
image = Image.open('68.jpeg') #path of the input image
enc_image = model.encode_image(image).to(device)

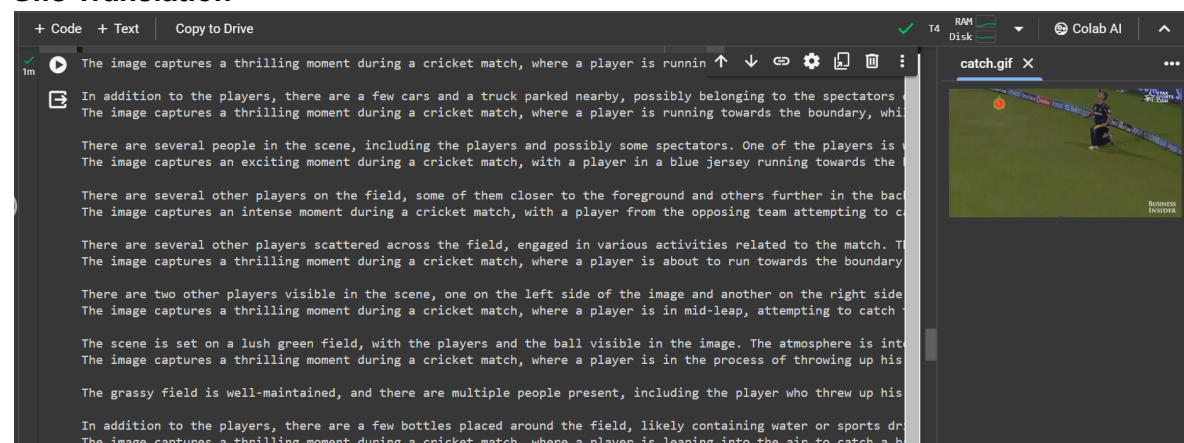
# Generate the answer
answer = model.answer_question(enc_image, " why is this edible", tokenizer)

# If the output is a tensor, move it back to CPU for further operations like print
if isinstance(answer, torch.Tensor):
    answer = answer.cpu()

print(answer)
```

The man in the image is a professional cricketer, and he is taking a swing at the ball with his bat.

Gifs Translation



The screenshot displays a Google Colab notebook interface. The left pane shows a list of generated text descriptions for a GIF. The right pane shows the GIF of a cricket player in a blue jersey running towards the boundary.

The image captures a thrilling moment during a cricket match, where a player is running towards the boundary, while the other players are attempting to catch the ball.

In addition to the players, there are a few cars and a truck parked nearby, possibly belonging to the spectators.

The image captures a thrilling moment during a cricket match, where a player is running towards the boundary, while the other players are attempting to catch the ball.

There are several people in the scene, including the players and possibly some spectators. One of the players is in a blue jersey, and the other players are in green jerseys.

The image captures an exciting moment during a cricket match, with a player in a blue jersey running towards the boundary, while the other players are attempting to catch the ball.

There are several other players on the field, some of them closer to the foreground and others further in the background.

The image captures an intense moment during a cricket match, with a player from the opposing team attempting to catch the ball.

There are several other players scattered across the field, engaged in various activities related to the match. The player in the blue jersey is running towards the boundary, while the other players are attempting to catch the ball.

The image captures a thrilling moment during a cricket match, where a player is about to run towards the boundary, while the other players are attempting to catch the ball.

There are two other players visible in the scene, one on the left side of the image and another on the right side.

The image captures a thrilling moment during a cricket match, where a player is in mid-leap, attempting to catch the ball.

The scene is set on a lush green field, with the players and the ball visible in the image. The atmosphere is intense, and the players are fully engaged in the match.

The image captures a thrilling moment during a cricket match, where a player is in the process of throwing up his hands in celebration.

The grassy field is well-maintained, and there are multiple people present, including the player who threw up his hands.

In addition to the players, there are a few bottles placed around the field, likely containing water or sports drinks.

The image captures a thrilling moment during a cricket match, where a player is leaping into the air to catch a ball.