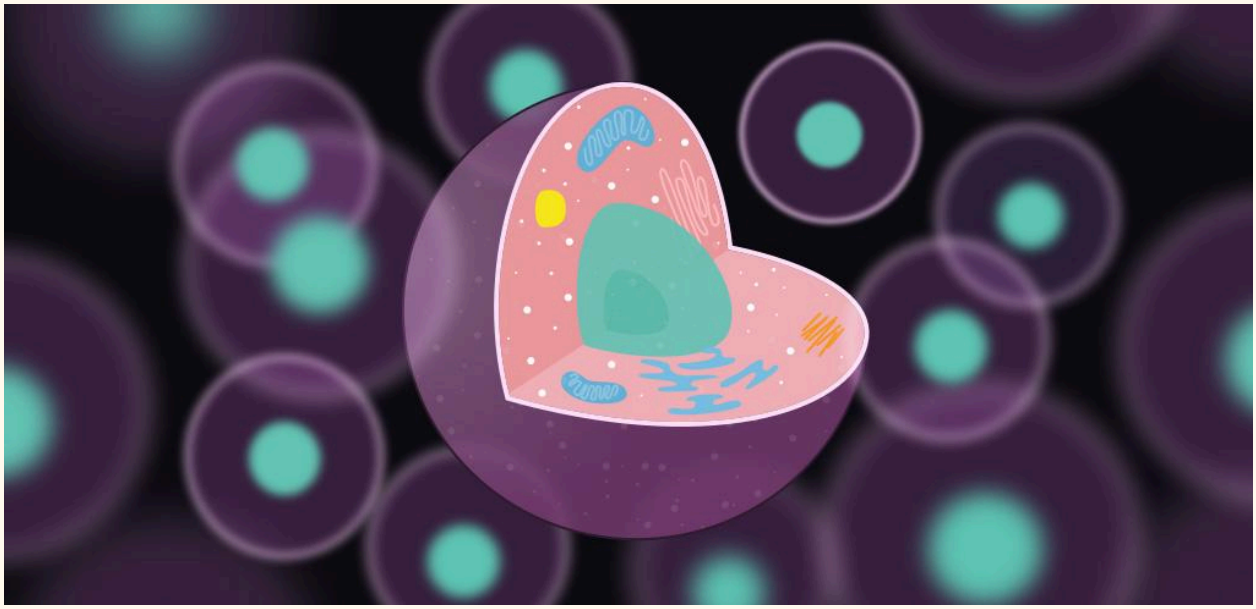
VISUAL QUESTION ANSWERING

By Sofia Raheel



INTRODUCTION

This project implements a web-based Visual Question Answering (VQA) system using the BLIP model. Users can upload an image and ask a relevant question in natural language. The application processes the image and question through the BLIP model to generate an accurate, context-aware answer. Built with Python, PyTorch, and Gradio for the interface, the project showcases the practical application of multimodal AI in real-time interactive settings.n.

BLIP

BLIP (Bootstrapping Language-Image Pre-training) is a vision-language model developed to bridge the gap between visual understanding and natural language processing. It leverages large-scale pre-training on image-text pairs to perform a variety of tasks such as image captioning, visual question answering (VQA), and image-text retrieval. The model is designed

with a flexible architecture that combines vision transformers (ViTs) with language models, enabling it to generate meaningful language outputs conditioned on visual inputs.

What makes BLIP particularly effective is its ability to handle both zero-shot and fine-tuned scenarios with high accuracy. It uses a novel captioning-based pre-training approach, allowing it to generalize well across different visual-language tasks. By aligning vision and language representations during training, BLIP becomes capable of understanding the context of an image and producing coherent textual responses to related queries.

Code Explanation

The codebase centers around using the BLIP model for Visual Question Answering (VQA) and consists of the following main components:

- **`app.py`**: This is the entry point of the application. It loads the pre-trained BLIP model and handles the logic for processing images and questions. It uses the `BlipForQuestionAnswering` model and a `BlipProcessor` from Hugging Face to perform inference. When a user submits an image and a question, the model generates a relevant answer based on the visual and textual input.

- **Model Loading & Inference**: The BLIP model (`Salesforce/blip-vqa-base`) is loaded with `from_pretrained`, along with its processor. The processor tokenizes the input question and prepares the image for the model. The output is then decoded to generate a readable answer using the model's `generate` method.

- **Gradio Interface**: The user interface is built using Gradio, allowing users to interact with the model through a simple web UI. Users upload an image and enter a question, and the app displays the model's answer in real time. The `gr.Interface` ties the input fields to the prediction function defined in `app.py`.

- **Requirements**: The project uses `transformers`, `torch`, `gradio`, and `Pillow` for model handling, computation, web interface, and image processing respectively. All dependencies are listed in `requirements.txt`.