Project write up:

Name:Hasbi Fathima VP

Enroll: E22CSEU0750

VisionIQ – A Visual Question Answering (VQA) App

VisionIQ is an advanced Visual Question Answering (VQA) project that combines state-of-the-art deep learning models to enable intelligent question-answering based on image inputs. VQA is a complex task that integrates **computer vision** and **natural language processing (NLP)**, requiring the model to understand both visual and textual data. VisionIQ leverages a combination of **Convolutional Neural Networks (CNNs)** for visual feature extraction and **Long Short-Term Memory (LSTM)** networks for sequential data processing. Additionally, it integrates a transformer-based **BLIP (Bootstrapping Language-Image Pre-training)** model from the HuggingFace library, allowing for enhanced alignment between visual and language representations.

Introduction

VQA is a challenging AI task that involves answering questions based on image inputs. Unlike text-based question-answering systems, VQA requires a deep understanding of both the image and the associated question. VisionIQ addresses this challenge by combining CNN-based visual feature extraction and LSTM-based sequence generation, enhanced by transformer-based multimodal alignment using the BLIP model. The project aims to create a powerful tool that generates accurate answers based on complex image-text interactions.

Technical Stack

- **Python** – Core language for model development and deployment.
- **Streamlit** – Framework for creating an interactive frontend interface.
- **HuggingFace Transformers** – BLIP model for vision-language alignment.
- **PyTorch** – For model training and inference.
- **OpenCV** – For image preprocessing and augmentation.
- **CNN** – For extracting visual features from images.
- **LSTM** – For processing sequential data and generating responses.

Training Strategy

- Dataset: Pretrained on **TextVQA** and other multimodal datasets from HuggingFace.
- Loss: Cross-entropy loss for training answer generation.
- Optimization: Learning rate scheduling and early stopping for performance improvement.

Deployment

- **Frontend:** Built with Streamlit for real-time user interaction.
- **Backend:** Uses FastAPI for model inference and API handling.

Use Cases

- **Education:** Assisting students with visual content-based learning.
- **Healthcare:** Interpreting medical images with text-based answers.
- **E-commerce:** Generating product descriptions from images.

Conclusion

VisionIQ demonstrates the power of combining CNN, LSTM, and transformer-based models to solve complex VQA tasks. The integration of BLIP for multimodal fusion enhances the model's ability to generate accurate and context-aware answers, making it a valuable tool across various industries.