Project Report: Automated Image Captioning System

1. Introduction

With the rapid advancement of AI, image captioning has become a crucial application in computer vision and natural language processing. This project focuses on implementing an automated image captioning system using **BLIP (Bootstrapped Language-Image Pretraining)** with a **FastAPI backend** and a **React frontend**.

2. Objective

The primary goal of this project is to develop an AI-powered system that can generate **detailed and accurate descriptions** for uploaded images. The system is designed to enhance accessibility, automate metadata generation, and improve content organization.

3. System Overview

3.1 Architecture

The system is structured into three main components:

- **Frontend:** A React-based interface for users to upload images.
- **Backend:** A FastAPI server handling image processing and model inference.
- **AI Model:** A BLIP-based model (Salesforce/blip-image-captioning-large) for generating captions.

3.2 Workflow

1. The user uploads an image through the React frontend.
2. The image is sent to the FastAPI backend.
3. The backend processes the image and passes it to the BLIP model.
4. The model generates a caption, which is returned to the frontend.
5. The caption is displayed to the user.

4. Technologies Used

Component	Technology
Frontend	React, TypeScript, Axios
Backend	FastAPI, Python, Pydantic
AI Model	Hugging Face Transformers (BLIP)
Deployment	(To be hosted on AWS)

5. Implementation Details

5.1 Backend - FastAPI

- Handles API requests for image uploads.
- Integrates with the BLIP model for caption generation.

- Uses **PIL** for image preprocessing.

5.2 AI Model - BLIP

- Uses a Transformer-based approach for **vision-language understanding**.
- Generates detailed captions using **beam search** and **sampling**.
- Runs on **CUDA (GPU)** for faster inference.

5.3 Frontend - React

- Provides an intuitive UI for uploading images.
- Sends requests to the FastAPI backend.
- Displays generated captions dynamically.

6. Results & Performance

- The model generates **coherent and contextually rich** captions.
- The API processes images in **under 3 seconds** on a GPU-enabled machine.
- The system successfully handles **various image types and resolutions**.

7. Future Improvements

- Optimize model inference for **real-time caption generation**.
- Extend support for **multiple captioning models**.
- Deploy the system using **AWS/GCP** for scalability.
- Improve the UI/UX for a seamless user experience.

8. Conclusion

This project demonstrates the power of AI in **automated image understanding**. By integrating computer vision and NLP, the system can **enhance accessibility, streamline content management, and improve digital interactions**. Future enhancements will focus on **performance optimization and cloud deployment**.

9. References

- Salesforce BLIP Model: https://huggingface.co/Salesforce/blip-image-captioning-large
- FastAPI Documentation: https://fastapi.tiangolo.com/
- React.js: https://react.dev/
