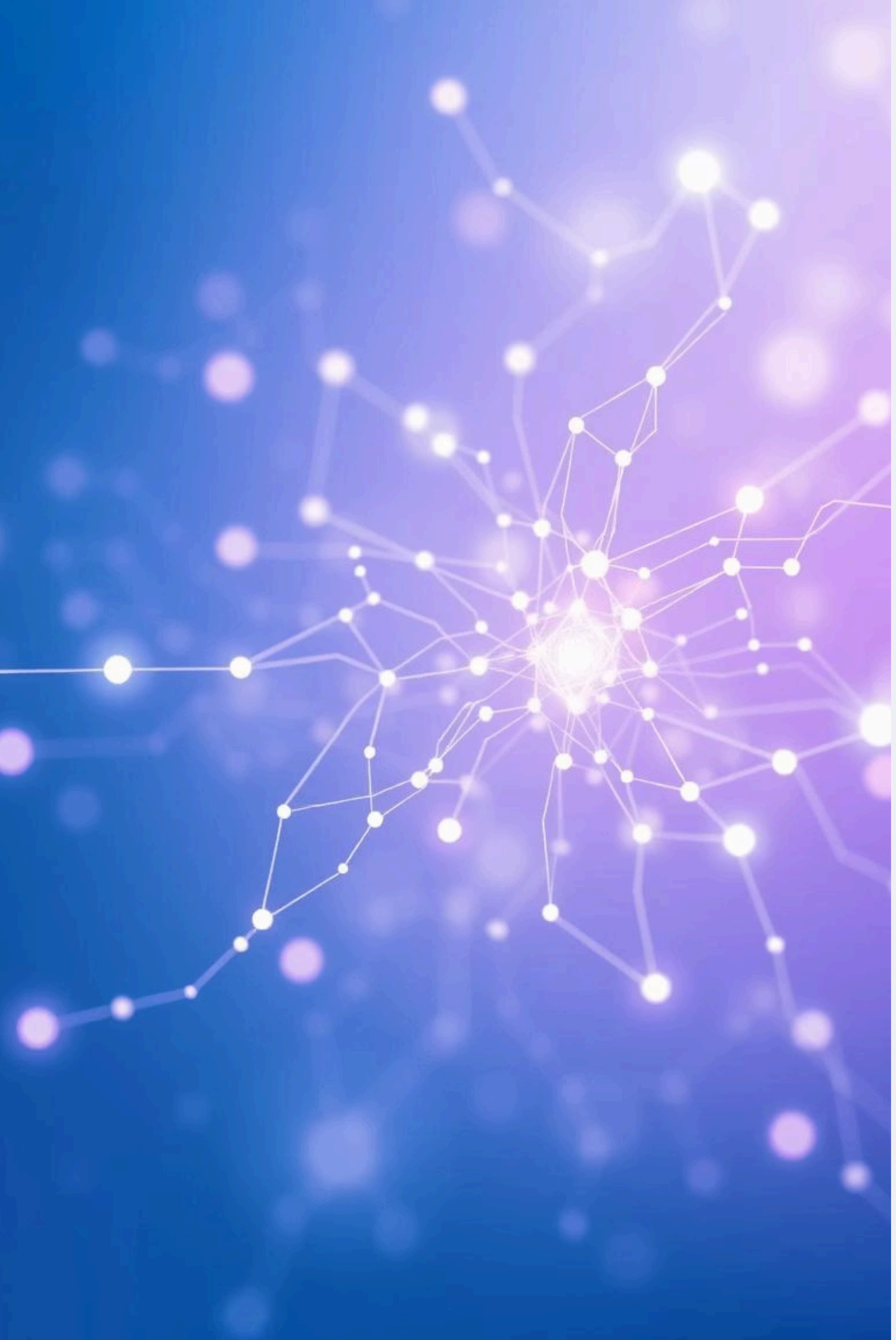


Vision.ai: AI-Powered Image Captioning for the Visually Impaired

- Vision.ai empowers visually impaired users through automated image captioning and speech output.
- The core team includes:
  - Nikhil Gupta (Lead Developer, model and backend)
  - Tanish Gupta and Harsit (frontend and integration)
  - Ajay Rajawat (testing)
- The tool leverages advanced machine learning and natural language processing to enhance accessibility and user experience.

Project Overview and Core Functionality

Core Purpose

Vision.ai assists visually impaired users by generating relevant image captions and converting those captions to audible speech, enabling enhanced interaction with visual media.

Technologies Used

- Image captioning with CNN and Transformer architectures
- Text-to-Speech (TTS) integration for audio output
- Optional caption translation using NLP models

Problem Statement: Addressing Accessibility Gaps

Target Audience

Individuals with visual impairments who struggle to interpret images independently.

Challenges

- Difficulty understanding visual content online
- Limited accessibility in current image viewing technologies

Solution

An integrated platform that automatically generates descriptive captions and reads them aloud, with optional translation features for broader usability.

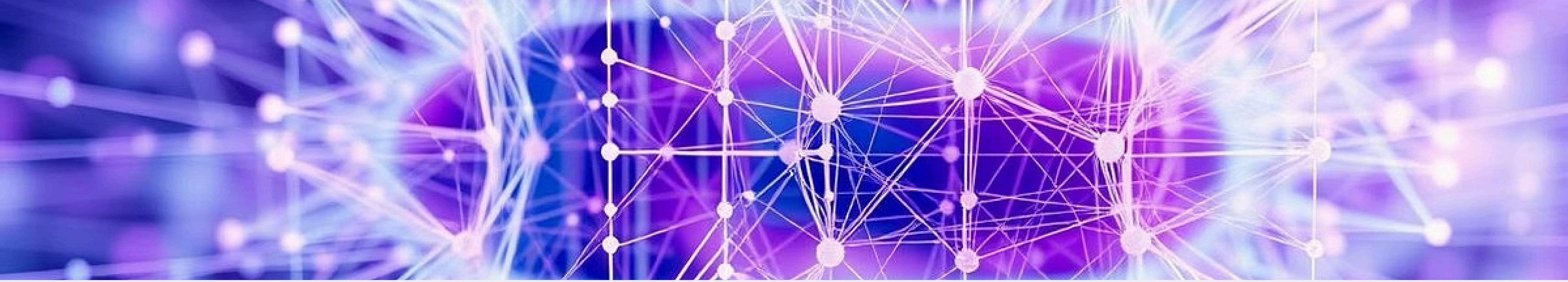
Architecture and Workflow of Vision.ai

Frontend: Streamlit

- Simple image upload interface
- Caption display and optional language selection
- User-friendly and accessible design

Backend: FastAPI

- Image processing and caption generation by CNN + Transformer model
- Translation via Helsinki-NLP models (optional)
- Sends captions and translations back to frontend



Technical Implementation Details

1

Model Development

Uses CNNs for image feature extraction and Transformers for sequence caption generation, training on the Flickr8k dataset for robust performance.

2

Translation Module

Integrates pre-trained Helsinki-NLP/opus-mt models enabling quick and accurate language translation of captions as per user preference.

3

Text-to-Speech Engine

Employs TTS tools such as Google Text-to-Speech (gTTS) to convert generated captions into clear, audible speech for users.

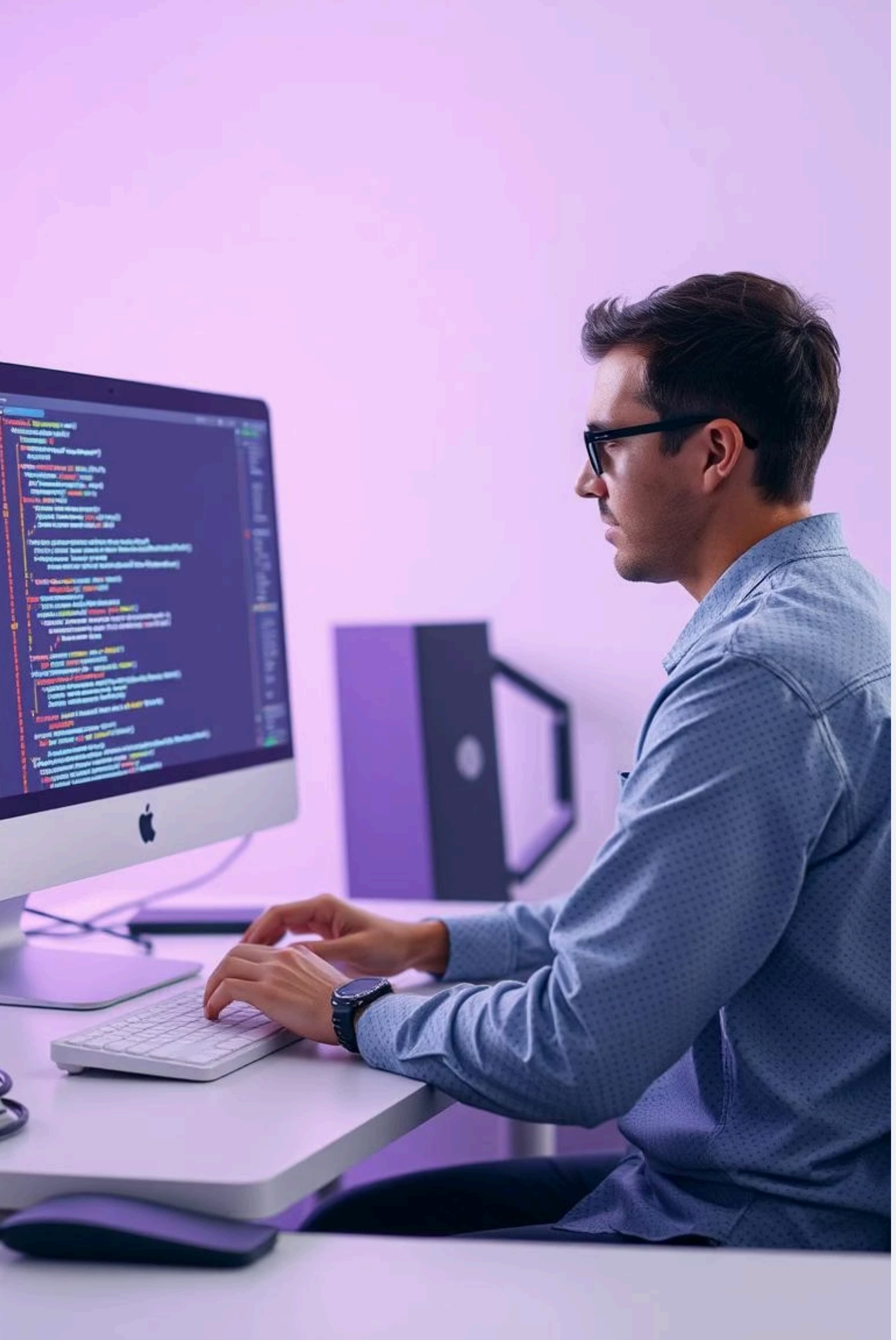
Deployment Strategy and Scalability

Current State

Locally deployed backend with FastAPI and frontend powered by Streamlit, communicating via HTTP requests for rapid prototyping and testing.

Future Plans

- Containerize backend for deployment on cloud platforms like AWS, Google Cloud or Azure
- Deploy frontend via Streamlit Community Cloud or dockerized environments for reliability and scalability



Key Challenges and Solutions

Real-Time Caption Generation

Optimized models and efficient server resources including GPUs allow rapid image processing and captioning without delay.

Accurate Multilingual Translation

Helsinki-NLP pretrained models provide fast, reliable translations maintaining caption meaning across diverse languages.

Accessibility and UX

Intuitive interface design combined with seamless Text-to-Speech functionality ensures ease of use for visually impaired users.

Testing, QA, and Project Impact

1

Testing Framework

Pytest utilized for unit and integration testing ensuring functionality of image captioning, API endpoints, and TTS outputs.

2

Project Impact

Vision.ai enables visually impaired individuals to access and comprehend image content, enhancing independence and digital inclusion.

3

Future Enhancements

Plans include model optimization, expanded language support, and added accessibility features such as speech recognition integration.