Automated Image Captioning and Speech Synthesis

INTERNSHIP REPORT

Submitted in partial fulfillment of the
Requirement for the award of the
Degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE ENGINEERING

By

KOMPALLY AKSHATHA
(2451-21-733-001)
KANDUKURI AKHIL KUMAR
(2451-21-733-022)

Under the guidance of
Dr. Sudha Pelluri
Head of the Department of CSE,
College of Engineering, Osmania University

Maturi Venkata Subba Rao Engineering College
Nadergul, Saroornagar Mandal, Hyderabad,Telangana-501510

May 2024 - July 2024

Automated Image Captioning and Speech Synthesis

INTERNSHIP REPORT

Submitted in partial fulfillment of the
Requirement for the award of the
Degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE ENGINEERING

By

KOMPALLY AKSHATHA
(2451-21-733-001)
KANDUKURI AKHIL KUMAR
(2451-21-733-022)

Under the guidance of
Dr. Sudha Pelluri
Head of the Department of CSE,
College of Engineering, Osmania University

COLLEGE OF ENGINEERING, OSMANIA UNIVERSITY
Department of Computer Science and Engineering
Amberpet, Hyderabad, Telangana 500007

May 2024 - July 2024

COLLEGE OF ENGINEERING, OSMANIA UNIVERSITY
Department of Computer Science and Engineering
Amberpet, Hyderabad, Telangana 500007

This is to certify that Summer Training entitled "**Automated Image Captioning and Speech Synthesis**" is a bonafide work carried out by **Kompally Akshatha (2451-21-733-001)** in partial fulfillment of the requirements for the award of degree of **Bachelor of Engineering** in **Computer Science and Engineering** from **Maturi Venkata Subba Rao(MVSR) Engineering College**, affiliated to Osmania University, Hyderabad during the Academic Year 2023-2024, under my guidance and supervision.

The results embodied in this report have not been submitted to any other university or institute for the award of any degree or diploma.

Head of the Department Dr.Sudha
Pelluri Professor & Head
Department of CSE College of
Engineering, Osmania University

COLLEGE OF ENGINEERING, OSMANIA UNIVERSITY
Department of Computer Science and Engineering
Amberpet, Hyderabad, Telangana 500007

This is to certify that Summer Training entitled "**Automated Image Captioning and Speech Synthesis**" is a bonafide work carried out by **Kandukuri Akhil Kumar (2451-21-733-022)** in partial fulfillment of the requirements for the award of degree of **Bachelor of Engineering** in **Computer Science and Engineering** from **Maturi Venkata Subba Rao(MVSR) Engineering College**, affiliated to Osmania University, Hyderabad during the Academic Year 2023-2024, under my guidance and supervision.

The results embodied in this report have not been submitted to any other university or institute for the award of any degree or diploma.

Head of the Department Dr.Sudha
Pelluri Professor & Head
Department of CSE College of
Engineering, Osmania University

DECLARATION

This is to certify that the work reported in the present project entitled "**Automated Image Captioning and Speech Synthesis**" is a record of bonafide work done by us in the Department of Computer Science and Engineering, MVSR Engineering College, Osmania University. The reports are based on the work done entirely by us and not copied from any other source.

The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma to the best of our knowledge and belief

Student
Kompally Akshatha
(2451-21-733-001)

DECLARATION

This is to certify that the work reported in the present project entitled "**Automated Image Captioning and Speech Synthesis**" is a record of bonafide work done by us in the Department of Computer Science and Engineering, MVSR Engineering College, Osmania University. The reports are based on the work done entirely by us and not copied from any other source.

The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma to the best of our knowledge and belief

Student
Kandukuri Akhil Kumar
(2451-21-733-022)

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Dr. Sudha Pelluri**, Head of the Department of Computer Science and Engineering at the College of Engineering, Osmania University, for her invaluable supervision and guidance during my internship. Her expertise and encouragement were instrumental in the successful completion of this project.

I am also deeply thankful to Osmania University for providing me with the opportunity to work on this project, which has significantly enhanced my knowledge and skills in the field of computer vision. The experience gained from this internship has been immensely beneficial and has provided me with practical insights into the application of advanced technologies

SUBMITTED BY
KOMPALLY AKSHATHA
KANDUKURI AKHIL KUMAR

ABSTRACT

Automated image captioning and speech synthesis are pivotal technologies intersecting computer vision and natural language processing. This study investigates the integration of Vision Transformer (ViT) models with Generative Pre-trained Transformers (GPT) for generating descriptive captions from images and synthesizing speech from these captions. The approach leverages ViT to extract high-dimensional visual features, subsequently fed into GPT to generate coherent, contextually relevant textual descriptions.

This project introduces an innovative web-based system combining automated image captioning and speech synthesis using cutting-edge AI technologies. Developed on the Flask framework, the platform integrates Vision Transformer (ViT) and Generative Pre-trained Transformer (GPT) models from the Transformers library. Users can upload images or capture them in real-time through the interface.

Upon image submission, the system per-processes images using OpenCV and PIL libraries to enhance quality and extract relevant visual features. These features are processed by a ViT model for high-dimensional feature extraction, followed by a GPT model to generate descriptive captions. The integration of ViT and GPT ensures accurate, contextually relevant captions across various image types and complexities.

To enhance accessibility, generated captions are transformed into synthesized speech using Google Text-to-Speech (gTTS), providing auditory feedback alongside visual captions. The platform offers a responsive user interface, an efficient image processing pipeline, and leverages GPU acceleration for enhanced performance.

Evaluation metrics include BLEU scores and qualitative assessments, confirming the system's efficacy in generating human-like captions and supporting diverse multimedia interaction needs.
This project exemplifies AI's application in improving user interaction through automated image understanding and accessible multimedia outputs.

.

TABLE OF CONTENT

1. INTRODUCTION

1.1 PROBLEM STATEMENT

In today's digital age, accessibility to information and content is paramount. For individuals with visual impairments, accessing and understanding visual content on the web poses significant challenges. Automated image captioning, combined with speech synthesis, offers a solution to bridge this gap, enabling visually impaired users to comprehend and engage with visual media effectively.

The primary challenge is to develop a web application that can automatically generate descriptive captions for images and synthesize these captions into speech. This solution should cater specifically to users who are blind or have low vision, allowing them to upload or capture images and receive meaningful descriptions through audio output. The system needs to be intuitive, accurate, and efficient, ensuring that users can easily interact with it using voice commands.

1.2 OBJECTIVES

- **To ensure accuracy and precision:** Our goal is to develop algorithms that can generate descriptive captions from images and synthesize natural-sounding speech from these captions.
- **To prioritize accessibility:** Our aim is to create a system that enhances inclusivity by providing auditory feedback alongside visual captions, catering specifically to users with visual impairments.
- **To enhance user experience:** Our focus is on improving interaction with multimedia content through seamless integration of visual and auditory modalities, thereby enhancing engagement and usability.
- **To optimize efficiency:** Our objective is to automate the processes of caption generation and speech synthesis, streamlining content creation and dissemination for faster and more efficient workflows.
- **To leverage technological advancement:** Our strategy involves utilizing cutting-edge AI technologies such as Vision Transformers and Generative Pre-trained Transformers to advance capabilities in image understanding and natural language processing.
- **To support research and development:** Our goal is to provide a platform for experimentation and advancement of algorithms in fields like computer vision and speech synthesis, contributing to ongoing AI research.
- **To enhance commercial viability:** Our objective is to integrate automated image captioning and speech synthesis capabilities, thereby enhancing the appeal and accessibility of digital products and services.

1.3 MOTIVATIONS

1. Ensure equal access to digital content for individuals with disabilities, promoting inclusivity and independence in accessing information.

2. Enhance user satisfaction and engagement by providing interactive and accessible multimedia experiences that cater to diverse user preferences.

3. Drive technological advancement by integrating advanced AI models and algorithms, pushing the boundaries of image understanding and speech synthesis capabilities.

4. Optimize work flows and increase productivity in content creation and dissemination through AI-driven automation, reducing manual effort and operational costs.

5. Foster educational and research initiatives by providing tools and platforms that facilitate experimentation and innovation in AI-driven multimedia processing and synthesis.

1.4 SCOPE

This project aims to develop a sophisticated web-based platform for automated image captioning and speech synthesis, leveraging state-of-the-art AI technologies. The core functionality will include implementing deep learning models such as Transformers for image caption generation and Tacotron for converting these captions into natural-sounding speech. Users will interact with the platform through a user-friendly interface that supports uploading images, capturing images via camera input, and providing voice commands for both actions and speech synthesis prompts.

The platform will be built using Flask for backend development, integrating CV2 (OpenCV) for image processing tasks, Torch (PyTorch) for deep learning model deployment, and PIL (Python Imaging Library) for image manipulation. Accessibility will be a key feature, with support for voice input enabling users with visual impairments to interact seamlessly with the system. The project will prioritize accuracy in image captioning and naturalness in speech synthesis, ensuring that generated captions are contextually relevant and speech output is fluent and expressive.

Future enhancements may include real-time image processing capabilities, integration with persistent data storage for user preferences and uploaded content, user authentication for personalized experiences, and multilingual support for speech synthesis. The development process will adhere to user-centric design principles to ensure an intuitive and inclusive platform experience across various user demographics and application scenarios.

Overall, this project aims to deliver a comprehensive solution that meets the growing demand for accessible, accurate, and user-friendly automated image captioning and speech synthesis tools in both personal and professional settings.

2. SYSTEM REQUIREMENT SPECIFICATIONS

2.1 SOFTWARE REQUIREMENTS

Web Development Framework:

Flask: A Python-based micro web framework that's ideal for building web applications. It offers simplicity and flexibility for developing APIs and serving user interfaces.

Automated Image Captioning and Speech Synthesis:

Vision Transformer (ViT): A transformer-based model specifically designed for image classification tasks by treating images as sequences of patches. This can be used for extracting features from images before feeding them into language models.

GPT-2: A large-scale language model capable of generating human-like text based on the input provided. When combined with ViT, it can process image features and generate descriptive captions.

gTTS (Google Text-to-Speech): Library for converting text generated by the Vit GPT-2 model into spoken audio.

Supporting Libraries:

PyTorch: Deep learning library used for training and deploying neural networks, including ViT and GPT-2 implementations.

Transformers (Hugging Face): Library for working with transformer-based models like GPT-2, providing pre-trained models and utilities for fine-tuning and inference.

Flask: Used for building the web application backend and serving the user interface.

OpenCV: For image processing tasks within the Flask application.

Pillow(PIL): Use for image processing tasks such as loading and transforming images before feeding them into the ViT model.

2.2 HARDWARE COMPONENTS

- **CPU**: A multi-core processor (e.g., Intel Core i5 or higher) is essential for handling computational tasks involved in image processing, machine learning inference, and web server operations.

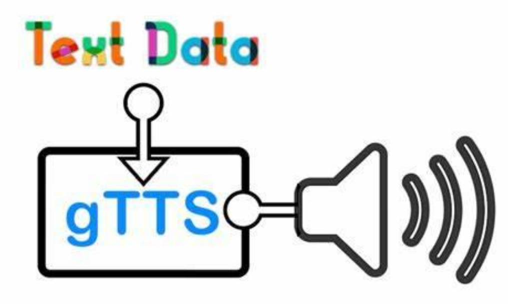
- **RAM**: Minimum 8 GB of RAM is recommended to ensure smooth operation of machine learning models and web server applications. More RAM (16 GB or higher) may be beneficial for handling larger datasets and concurrent user requests.

- **Storage**: SSD storage is preferable over HDD for faster data access and improved performance, especially when dealing with large image datasets and model checkpoints.

- **GPU**: GPU with CUDA support (e.g., NVIDIA GeForce GTX 1060 or higher) can significantly accelerate deep learning model training and inference tasks. For production-scale deployments, more powerful GPUs like NVIDIA Tesla V100 may be considered for faster processing speeds.

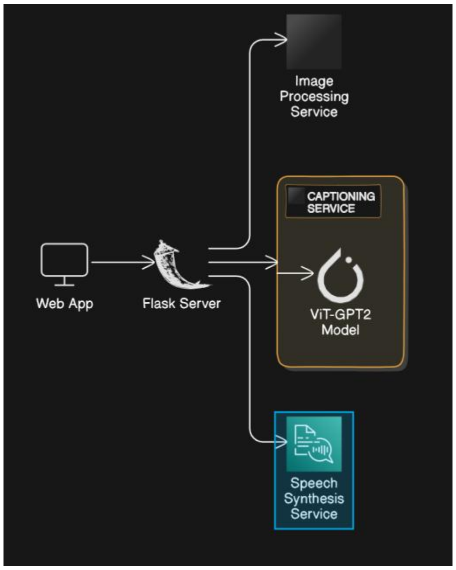
- **Network**: A stable internet connection is required for hosting the web application and handling user requests, especially if deploying on cloud services or remote servers.

These components will provide the necessary infrastructure to effectively run and scale your automated image captioning and speech synthesis application, ensuring efficient processing and responsiveness to user interactions.



3.SYSTEM DESIGN

3.1 SYSTEM ARCHITECTURE / BLOCK DIAGRAM



The diagram illustrates the architecture of an automated image captioning and speech synthesis website, encompassing several components that interact sequentially to deliver the final audio output. The process begins with a Web App, the user interface where users upload images. This user request is directed to the Flask Server, which functions as the backend server, orchestrating the communication between the web application and various backend services. The server forwards the uploaded images to the Image Processing Service, responsible for any necessary preprocessing tasks such as re sizing or normalization to prepare the images for captioning.

Following the preprocessing, the images are sent to the Captioning Service. At the heart of this service is the ViT-GPT2 model, a sophisticated deep learning model that integrates the Vision Transformer (ViT) for image analysis and GPT-2 for generating natural language descriptions. This model processes the images and produces descriptive captions. The generated captions are then passed to the Speech Synthesis Service, which converts the text captions into speech, creating an audio file.

The Flask Server receives this audio file and returns it to the user through the Web App. This entire process, from image upload to audio delivery, exemplifies a seamless integration of image processing, caption generation, and text-to-speech synthesis, providing an automated solution for converting images into spoken descriptions.

4.IMPLEMENTATION

4.1 ENVIRONMENTAL SETUP

- **Frameworks**: Flask or Django are popular choices for developing the backend of automated image captioning and speech synthesis applications due to their robustness and scalability.

- **Programming Languages**: Python is extensively used for its rich libraries in machine learning and web development, making it ideal for implementing automated image captioning and speech synthesis systems.

- **Version Control**: Git is essential for managing collaborative development and version control of the codebase across teams working on automated image captioning and speech synthesis projects.

- **Development Environment**: Integrated Development Environments (IDEs) like PyCharm or Visual Studio Code provide comprehensive tools and debugging capabilities crucial for developing automated image captioning and speech synthesis applications efficiently.

- **Python Libraries**: Libraries such as torch, transformers, OpenCV, Pillow (PIL), and gTTS are indispensable for implementing machine learning models, image processing, and text-to-speech functionalities in automated image captioning and speech synthesis projects.

4.2 Model Architecture

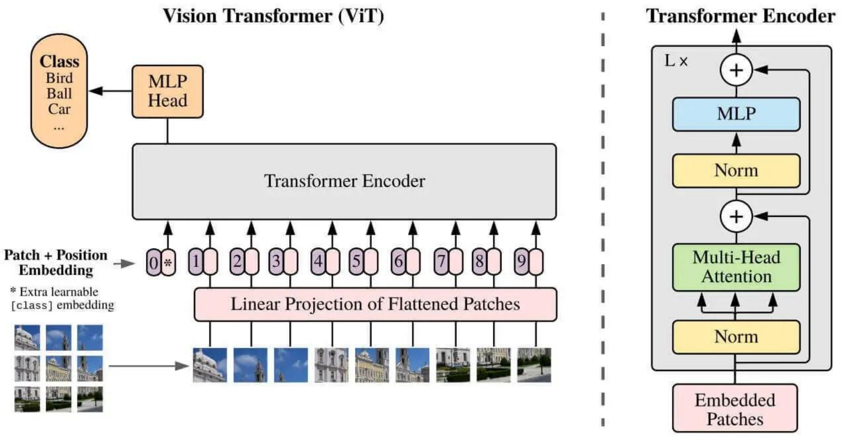


Fig:The Vision Transformer (ViT) architecture

The Vision Transformer (ViT) is a transformer-based model originally designed for image classification tasks, but it can also be adapted for tasks like image captioning. Here's a brief overview of the ViT architecture and its adaptation for image captioning:

Transformer Architecture: ViT applies the transformer architecture, which is widely used in natural language processing (NLP) tasks, to images. It consists of transformer blocks that process both spatial and positional information of image patches.

Input Representation: Unlike traditional convolutional neural networks (CNNs) that process images as grids of pixels, ViT splits the image into fixed-size patches and flattens each patch into a vector. These patch embeddings are then linearly embedded to form the input tokens for the transformer.

Positional Encoding: Since transformers do not inherently encode spatial information, ViT uses learned positional embeddings to inject spatial information about the patches' positions into the input tokens.

Transformer Encoder: The core of ViT consists of multiple transformer encoder blocks. Each block includes self-attention mechanisms to capture dependencies between different patches and feedforward neural networks for processing.

CLS Token: ViT uses a special learnable token called the "CLS" token, similar to its use in NLP transformers like BERT. This token aggregates information from the entire image and is used for classification tasks.

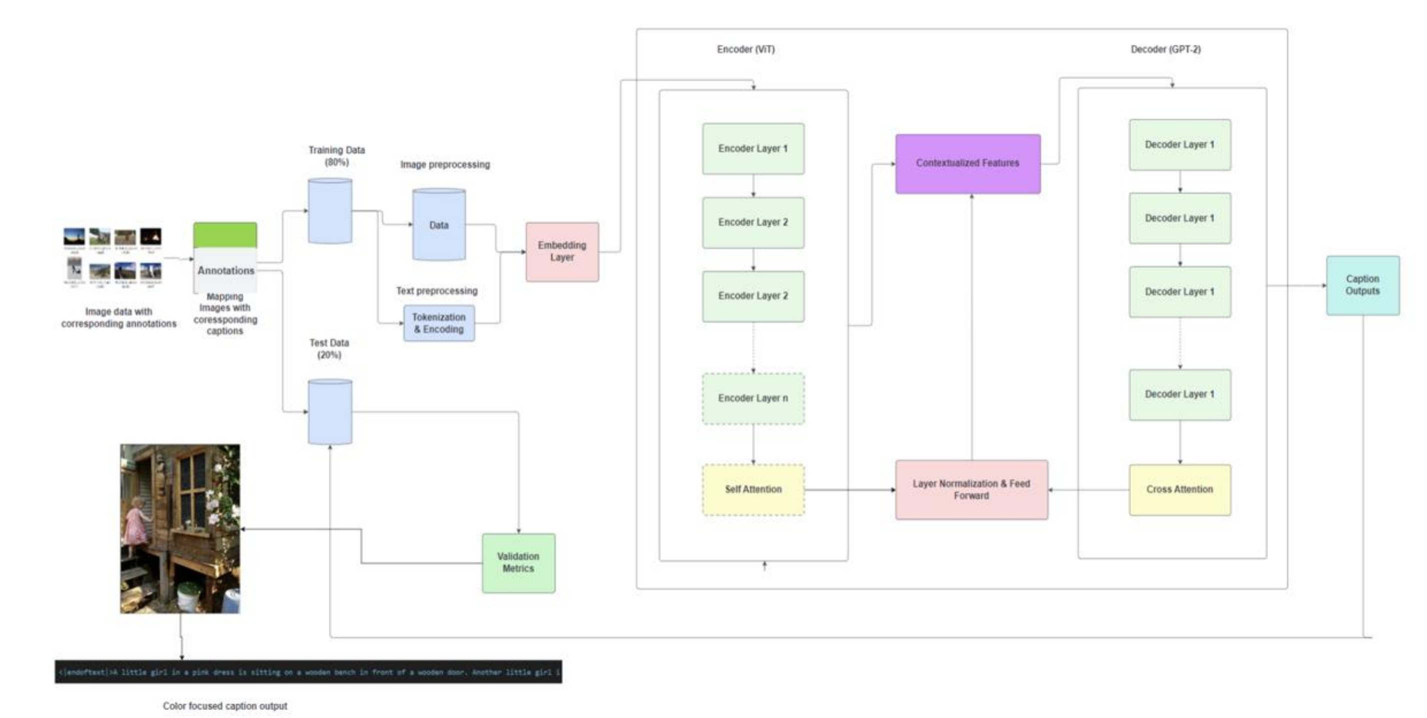
Adaptation for Image Captioning:

- **Encoding Images:** ViT can be adapted for image captioning by using the final hidden state of the CLS token or other tokens for context aggregation across patches.
- **Generating Captions:** Post-encoding, a decoder component can be added to generate captions. This decoder typically involves additional transformer layers or recurrent neural networks (RNNs) to produce sequences of words that describe the image content.

Training: ViT for image captioning requires pre-training on large-scale image datasets (e.g., ImageNet) followed by fine-tuning on captioning-specific datasets (e.g., COCO) to optimize performance for generating accurate and relevant captions.

Advantages: ViT offers several advantages such as scalability to larger image sizes, ability to capture long-range dependencies, and flexibility in handling various image-related tasks beyond classification, including captioning.

Adopting ViT for image captioning involves integrating its encoder with a suitable decoder architecture to handle the sequence generation aspect effectively, ensuring coherence and relevance in generated captions.



```
<|endoftext|>A little girl in a pink dress is sitting on a wooden bench in front of a wooden door. Another little girl i
```

4.2.1 Frontend Development

Home Page (home.html):

Purpose: This is the landing page of the web application, offering two primary Functionalities which are uploading an image or capturing an image using the webcam.

Design:

- The page features a visually appealing design with a linear gradient background.
- It contains a central container with two buttons: "Upload Image" and "Capture Image".
- The container has hover effects for interactivity, enhancing user engagement.

Upload Image Page (index.html):

Purpose: This page allows users to upload an image from their local device for captioning.

Design:

- A clean, responsive design with a linear gradient background.
- The page includes an upload form where users can select an image file and submit it.
- It has an animated container for the upload form, providing a smooth user experience.
- Additional UI elements like animated background and hover effects on the image and buttons.

Capture Image Page (index1.html):

Purpose: This page allows users to capture an image using their webcam for captioning.

Design:

- Similar visual style to the upload page, with a focus on webcam integration.
- Includes a video element displaying the webcam feed and a "Capture" button to take a snapshot.
- Captured images are automatically submitted for processing.

Result Page (result.html):

Purpose: This page displays the uploaded or captured image, along with its generated caption and synthesized speech.

Design:

- The page showcases the result with an image, its caption, and an audio player for the synthesized speech.
- A "Upload Another Image" button allows users to return to the home page and start the process again.
- Uses a visually engaging background and container design to maintain a consistent look and feel.

4.1.2 Backend Development

Framework:

- The backend is built using Flask, a lightweight Python web framework, which serves the web pages and handles requests.

Image Upload and Processing:

- **File Handling**: The backend processes uploaded images or captured images by saving them to a specified directory.
- **Preprocessing**: Uploaded images undergo preprocessing steps such as converting to RGB, enhancing grayscale and LAB components to improve the quality.

Image Captioning:

- **Model**: Uses a pre-trained VisionEncoderDecoderModel (ViT-GPT2) from the transformers library.
- **Feature Extraction**: Employs the ViTFeatureExtractor for extracting features from images.
- **Tokenization**: Uses the AutoTokenizer for decoding the generated captions.
- **Device Management**: The model is loaded onto a GPU if available, otherwise, it defaults to the CPU.

Caption Generation:

- The backend generates captions for the preprocessed images by passing them through the model.
- **Inference**: The model performs inference to generate textual descriptions of the images.

Speech Synthesis:

- **Text-to-Speech**: Utilizes the gTTS (Google Text-to-Speech) library to convert the generated captions into speech.
- The synthesized speech is saved as an audio file and linked to the result page.

Routes and Handlers:

- **Home Route**: Renders the home page.
- **Upload Image Route**: Renders the upload image page.
- **Capture Image Route**: Renders the capture image page.
- **Upload Handler**: Handles the form submission for uploaded images, processes the image, generates the caption, and redirects to the result page.
- **Capture Handler**: Handles the form submission for captured images similarly, processing them and generating the caption.

Result Presentation:

- **Dynamic Rendering**: The result page dynamically displays the uploaded/captured image, generated caption, and provides a playback option for the synthesized speech.
- **URL Management**: Ensures proper linking and retrieval of image and audio files for seamless user experience.

5.RESULTS

home.html

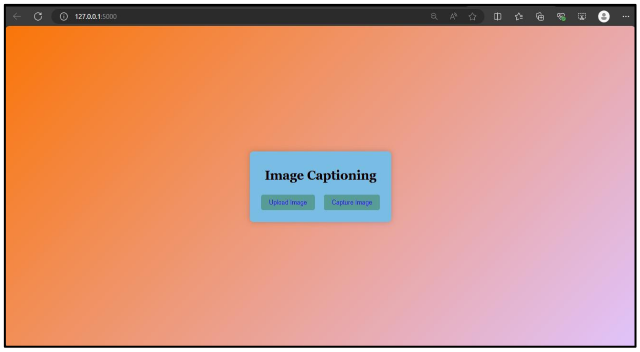
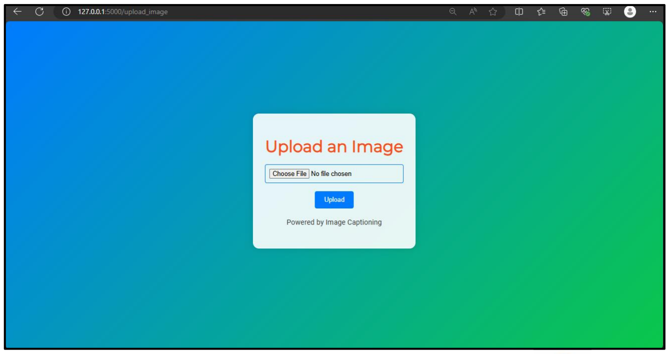


Image Captioning

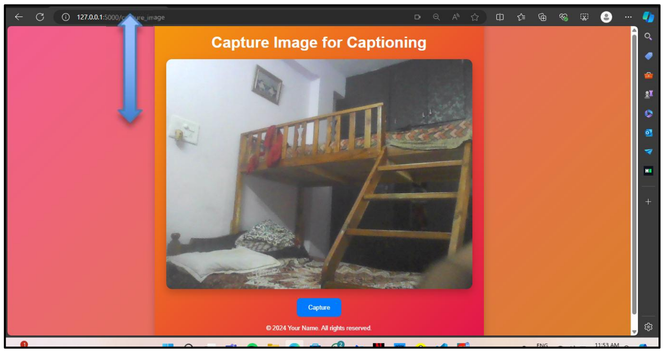
index.html



Upload an Image

Powered by Image Captioning

index1.html



Capture Image for Captioning

© 2024 Your Name. All rights reserved.

result.html

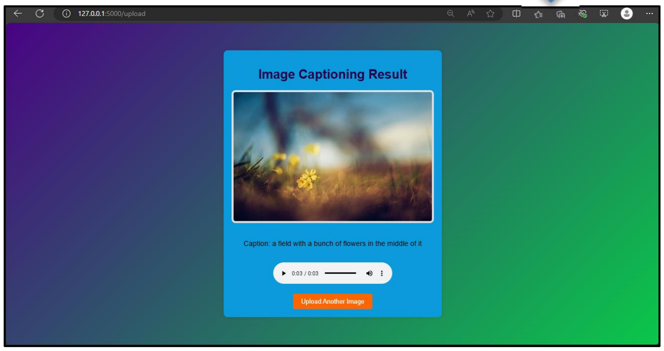


Image Captioning Result

Caption: a field with a bunch of flowers in the middle of it

Sample Output

result.html

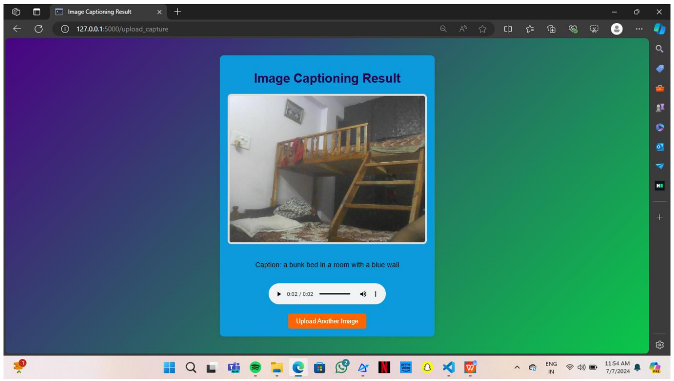


Image Captioning Result

Caption: a bunk bed in a room with a blue wall

Sample Output

7. CONCLUSION

In conclusion, developing an automated image captioning and speech synthesis website using the Vision Transformer (ViT) and GPT-2 model represents a significant advancement in multimedia processing and interaction technologies. This approach combines ViT's capability to extract detailed visual features from images with GPT-2's proficiency in generating coherent and contextually appropriate captions in natural language.

The integration of these technologies enables the website to automatically describe images with textual captions and further convert these captions into synthesized speech, enhancing accessibility for visually impaired users and improving user experience in multimedia content consumption.

Key advantages of this approach include:

- Accuracy and Contextual Understanding: ViT-GPT-2 integration ensures that generated captions accurately reflect the content and context of the images, providing meaningful descriptions that enhance understanding.

- Versatility and Adaptability: The website can handle various types of images and generate captions that are tailored to specific visual content, leveraging the broad training data and adaptive nature of GPT-2.

- Enhanced User Interaction: By enabling speech synthesis from generated captions, the website enhances user interaction by providing both visual and auditory access to image content.

Challenges such as computational complexity and ensuring alignment between visual and textual contexts should be addressed through optimization and fine-tuning strategies. Future advancements may focus on improving real-time processing capabilities and expanding the model's capability to handle diverse multimedia inputs.

Overall, an automated image captioning and speech synthesis website using ViT and GPT-2 represents a cutting-edge application of AI technologies, promising enhanced accessibility and user engagement in multimedia content consumption scenarios.

8.REFERENCES

1. Image Captioning by ViT/BERT, ViT/GPT https://www.researchgate.net/publication/369655809_Image_Captioning_by_ViTBERT_ViTGPT

2. Vision Encoder Decoder Models
https://github.com/Redcof/vit-gpt2-image-captioning
https://huggingface.co/docs/transformers/model_doc/vision-encoder-decoder

3. Multilingual & Color-Focused Image Captioning For Visually Impaired Using Deep Learning Techniques by Poojan Gagrani . Department of Applied Data Science, San José State University . DATA 270: Data Analytics Processes , Dr. Eduardo Chan , December 6, 2023

4. Image Captioning Using Hugging Face Vision Encoder Decoder — Step 2 Step Guide (Part 2) https://medium.com/@kalpeshmulye/image-captioning-using-huggingface-vision-encoder-decoder-step-2-step-guide-part-2-95f64f6b73b9

9. APPENDIX

SOURCE CODE:

Refer to the source code at: https://github.com/kakshatha-001/Auto_Img_Capt