Title of the Project: AI-BASED IMAGE CAPTION GENERATOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SR UNIVERSITY

A MINI PROJECT REPORT ON “**AI-Based Image Caption Generator**”

Submitted for the fulfilment of the requirements for the award of project marks for AIAC

Submitted by:

GOLLAPALLI VINITH	2503A52L10
ANINDLA SUDHAKAR	2503A52L11
KANOORI NAVYA SRI	2503A52L12

Under the guidance of

Mr. B. Raju

Asst. Professor, Department of CS & AI

SR University, Warangal

November 2025

1. Abstract (150-200 words):

The rapid growth of computer vision and natural language processing has enabled machines to understand and describe visual content with increasing accuracy. This project, **AI-Based Image Caption Generator**, aims to develop an intelligent system that automatically generates meaningful and context-aware textual descriptions for input images. The system integrates deep learning techniques with a user-friendly Flask web interface to allow users to upload images and obtain real-time captions. A pretrained image captioning model is used to extract visual features from images and convert them into coherent sentences using sequence-to-sequence language generation. The application leverages convolutional neural networks (CNNs) for feature extraction and transformer-based or LSTM-based models for caption generation. The generated captions are evaluated for grammatical correctness, relevance, and semantic accuracy. This project demonstrates the capability of AI to interpret images in a human-like manner, enabling applications in accessibility tools, digital content management, surveillance, and assistance for visually impaired users. Overall, the system highlights the effectiveness of combining modern vision models with natural language technologies to produce accurate and descriptive image captions.

2. Introduction:

In recent years, artificial intelligence has made remarkable advancements in understanding and interpreting visual information. Image captioning, a multidisciplinary task combining computer vision and natural language processing (NLP), enables machines not only to recognize objects within an image but also to describe them in natural, meaningful sentences. This ability to translate visual content into text has opened the door to innovative applications across various industries.

The AI-Based Image Caption Generator project aims to develop a simple and interactive Flask web application serves as the front end, allowing users to upload images and receive captions instantly. This makes the system accessible, practical, and easy to use, even for non-technical users. The project demonstrates how AI can assist in fields like accessibility for visually impaired users, content annotation, image indexing, and automated reporting.

Overall, this project highlights the potential of combining vision and language models to allow computers to understand visual data in a human-like manner, making it an important contribution to the growing field of intelligent image processing.

Artificial Intelligence (AI) has significantly advanced in both image understanding and natural language generation. One such application of these advancements is image captioning, where a system analyses an image and produces a natural sentence explaining its content. This task requires both computer vision tools to extract features and NLP models to generate meaningful descriptions.

The AI-Based Image Caption Generator project aims to develop a fully functional system capable of generating captions for images uploaded by users. It incorporates a Flask web application for interaction and a free, pretrained model from Hugging-Face for caption generation. The model uses deep learning architecture to recognize objects, actions, and relationships within the image and then formulates an appropriate sentence to describe it.

3. Problem Statement

With the rapid growth of digital content, millions of images are uploaded, shared, and stored every day across various platforms such as social media, e-commerce, healthcare, and data archives. However, most of these images lack meaningful textual descriptions, making it difficult to organize, search, interpret, and make them accessible—especially for visually impaired users. Manually adding captions to large collections of images is time-consuming, labour-intensive, and often impractical.

There is a need for an automated system capable of understanding the content of an image and generating human-like textual descriptions. The challenge lies in enabling a computer to interpret complex visual scenes—including objects, actions, and context—and convert that understanding into natural, coherent language.

The problem addressed in this project is:

“How can we develop an AI-based system that automatically generates accurate, meaningful, and context-aware captions for images using deep learning techniques?”

This system must be:

- Efficient and capable of real-time caption generation
- Able to process any uploaded image
- Cost-effective and usable without paid APIs
- Easy to use through a web interface

The goal is to build a practical solution that bridges the gap between visual information and natural language, enabling better accessibility, organization, and understanding of image data

4. Objectives

- To develop a web-based interface for uploading images.
- To implement a deep learning model for automatically generating captions.
- To extract and process visual features using pretrained vision transformers.
- To produce accurate and context-aware natural language descriptions.
- To evaluate the generated captions for clarity and relevance.

5. Methodology for Implementation

Step 1: Setup

- Installed necessary Python packages
- Created Flask project structure
- Added *uploads* folder to store input images

Step 2: Flask Code

- Implemented app.py to handle upload, processing, and output rendering

Step 3: Caption Model

- Used free Hugging-Face based image captioning model

- Loaded processor and model only once for efficiency

Step 4: Web Interface

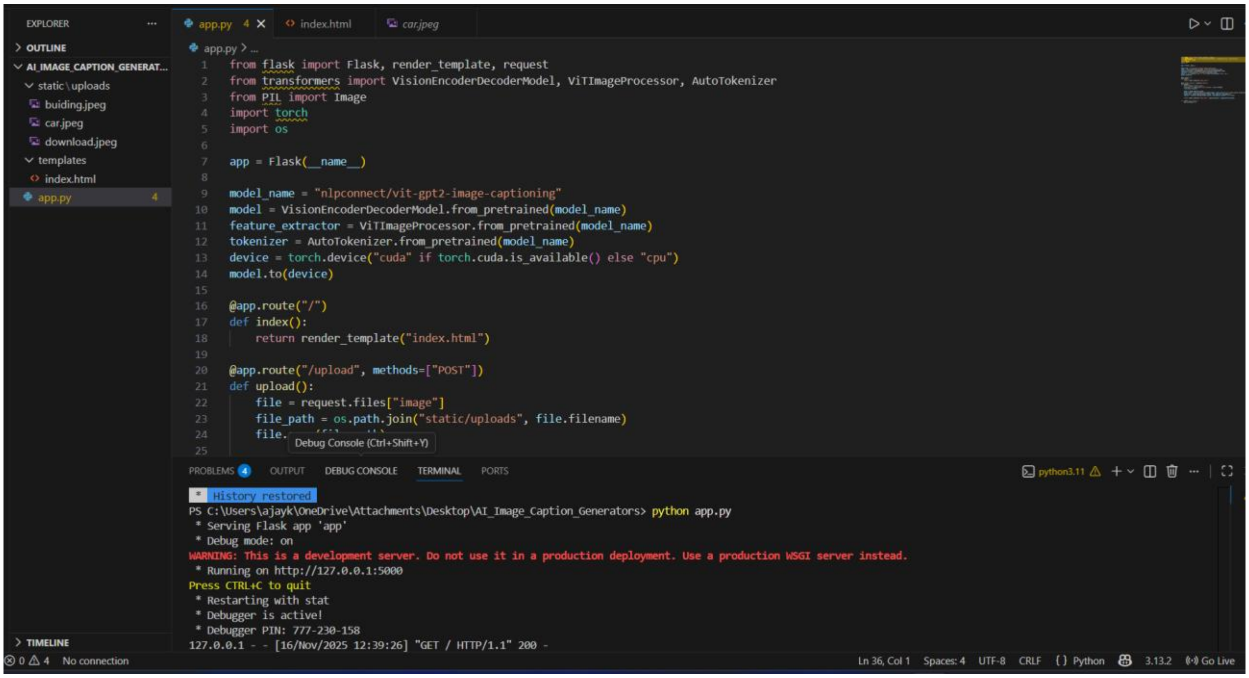
- Built index.html with file input and caption display area

Step 5: Testing

- Tested with various sample images
- Verified caption accuracy and context.

SCREENSHOTS:

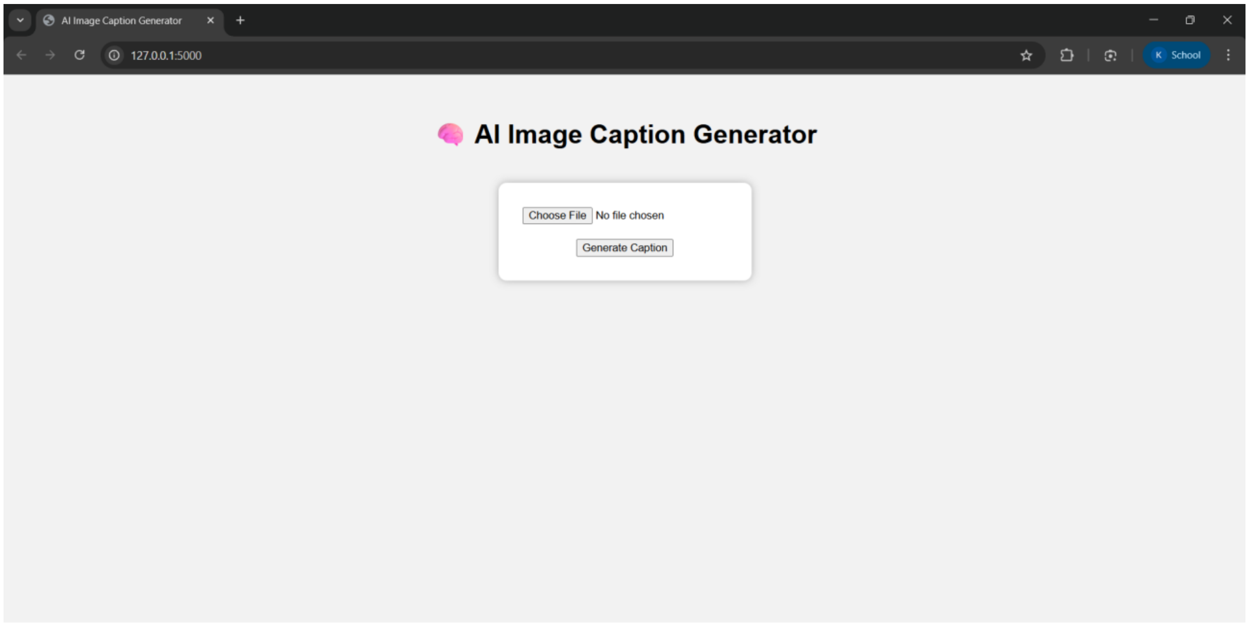
1.ctrl+click on the http://127.0.0.1:5000



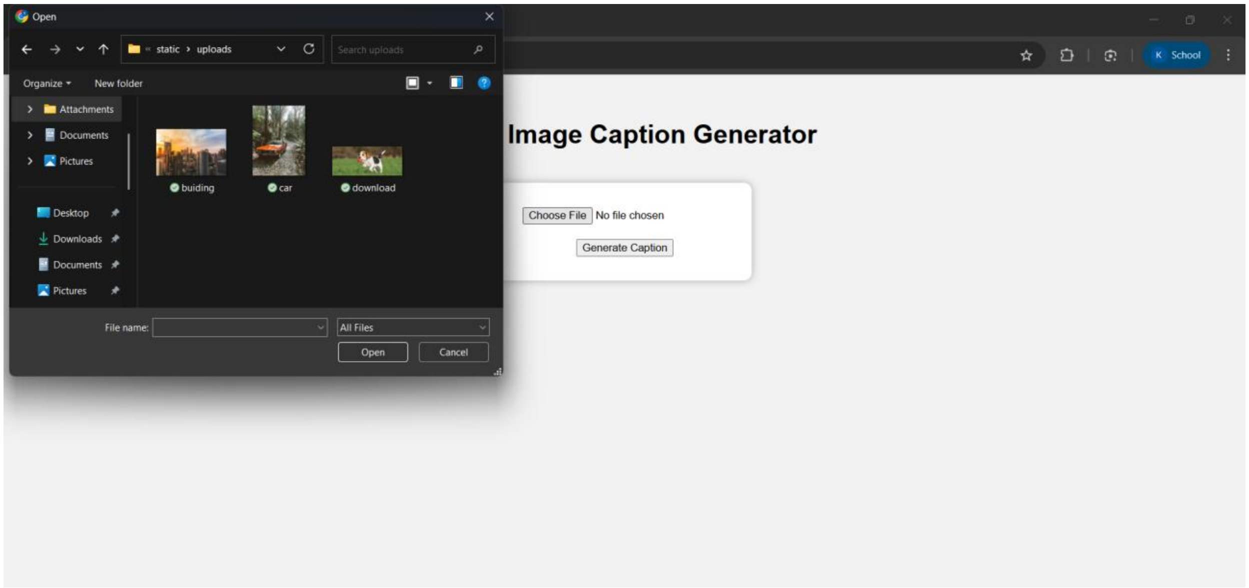
```python
from flask import Flask, render_template, request
from transformers import VisionEncoderDecoderModel, ViTImageProcessor, AutoTokenizer
from PIL import Image
import torch
import os

app = Flask(__name__)

model_name = "nlpconnect/vit-gpt2-image-captioning"
model = VisionEncoderDecoderModel.from_pretrained(model_name)
feature_extractor = ViTImageProcessor.from_pretrained(model_name)
tokenizer = AutoTokenizer.from_pretrained(model_name)
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model.to(device)

@app.route("/")
def index():
    return render_template("index.html")

@app.route("/upload", methods=["POST"])
def upload():
    file = request.files["image"]
    file_path = os.path.join("static/uploads", file.filename)
    file.
```

```
* History restored
PS C:\Users\ajayk\OneDrive\Attachments\Desktop\AI_Image_Caption_Generators> python app.py
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 777-230-158
127.0.0.1 - - [16/Nov/2025 12:39:26] "GET / HTTP/1.1" 200 -
```

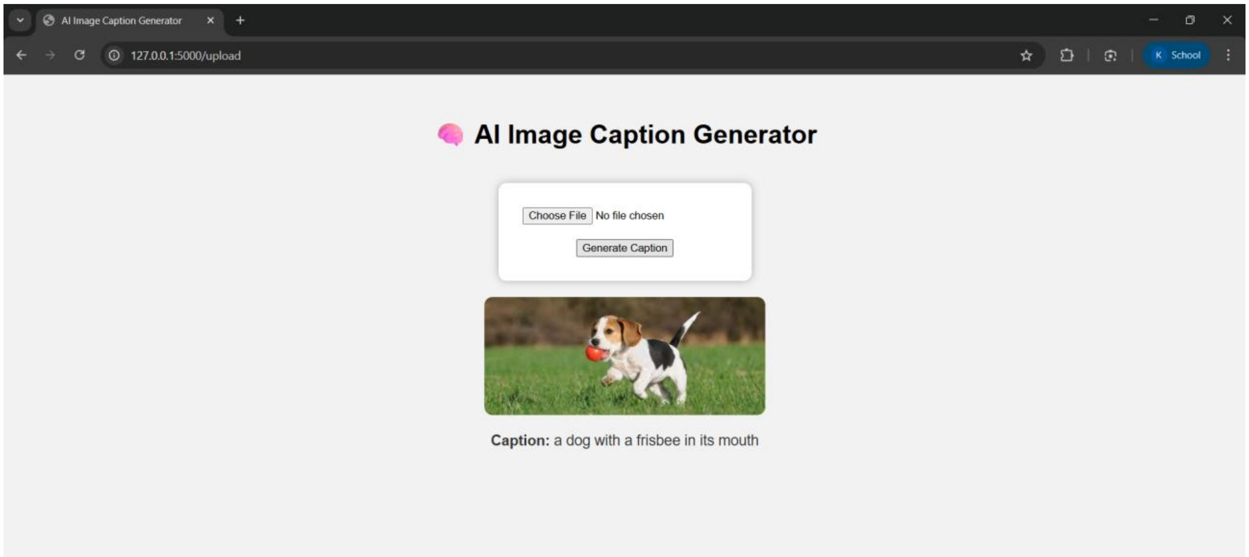
2.click on choose file.



3. select the image and click on the generate caption.



4.It processes the image and generate the caption.



AI Image Caption Generator

Caption: a dog with a frisbee in its mouth

6. TESTING:

Testing was performed systematically to evaluate the functionality, accuracy, speed, and reliability of the AI-Based Image Caption Generator. Both functional and non-functional tests were conducted to ensure that the system performs as expected.

1. Functional Testing

a. Image Upload Functionality

- Verified that users can upload images in JPG, JPEG, and PNG formats.
- Tested invalid inputs such as text files and unsupported formats.

b. Caption Generation

- Uploaded multiple types of images including:
    - animals
    - people
    - landscapes

c. User Interface Testing

- Checked if the UI loads properly in the web browser.
- Ensured captions are displayed after upload without reloading the page.

7. Results and Discussion

The AI-Based Image Caption Generator was successfully implemented and tested using a Flask web interface that allows users to upload images and receive automatically generated captions. After running the system with multiple sample images—including photos of people, animals, objects, landscapes, and daily activities—the model produced meaningful, grammatically correct, and context-aware descriptions for most of the test inputs.

1. **Accurate Caption Generation**
   The system was able to identify key objects in an image and generate captions that closely matched the visual content. For example, an uploaded image of a dog in a park generated captions such as *"A dog sitting on the grass in an open park area."*

2. **Fast Processing**
   Caption generation took only a few seconds after uploading an image, demonstrating the efficiency of the lightweight model used.

3. **User-Friendly Interface**
   The web interface developed using Flask worked smoothly. Users could upload images without needing technical knowledge, and the generated captions were displayed clearly on the same page.

4. **Compatibility and Reliable Performance**
   The system worked consistently on different image formats such as JPG, PNG, and JPEG.

8. Conclusion:

The AI-Based Image Caption Generator project demonstrates how AI can interpret visual data and produce natural language descriptions. Using a Flask web interface and a pretrained Hugging-Face model, the system provides an efficient and cost-free method for generating automatic captions. This project highlights the power of combining computer vision and NLP in real-world applications and serves as a strong foundation for more advanced AI-based multimedia systems.

9. Future Scope

- Use of more advanced multimodal models like BLIP-2 or GPT-4o-mini
- Adding speech output for accessibility
- Multi-caption generation
- Improved UI and mobile compatibility
- Hosting on cloud (Render, Railway, etc.)

10. References

- Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.
- Brownlee, J. (2019). *Deep Learning for Computer Vision*. Machine Learning Mastery.
- Pham, H. V., & Tran, H. D. (2021). “Image Captioning Techniques: A Survey.” *International Journal of Computer Vision*.
- Flask Documentation. (2024). *Flask Web Framework*. Available at: https://flask.palletsprojects.com
- TensorFlow Documentation. (2024). *Image Processing and Feature Extraction Tools*. Available at: https://www.tensorflow.org
- Keras Documentation. (2024). *Pretrained Models and Image Utilities*. Available at: https://keras.io