AUTOMATIC IMAGE CAPTION GENERATION USING DEEP LEARNING AND MACHINE LEARNING TECHNIQUES

[1]Mrs.B.Sowjanya ,[2]Reddy Anitha, [3]Palla Geethika, [4]Kola Sai Bhupathi Nitesh,[5]Paluri Sekhar, [6]T. Sasi Kiran,

[1]Assistant Professor, Computer Science andEngineering(AI -ML)

[2,3,4,5,6] Department of Computer Science and Engineering (Data Science)

[1,2,3,4,5,6]Avanthi Institute of Engineering and Technology (AVEN), Annakapalle, Andhra Pradesh,India.

[1]ssowjanyacse@gmail.com

[2]anithareddy3434@gmail.com

[3]kolabhupathi@gmail.com

[4]pallageethika@gmail.com

[5]palurisekhar@gmail.com

[6]tsasi1302@gmail.com

Abstract

Image captioning is an advanced application of Artificial Intelligence that combines Computer Vision and Natural Language Processing (NLP) to generate meaningful textual descriptions from images. With the rapid growth of digital media platforms, a massive amount of visual data is generated every day. Understanding and describing this visual content automatically is a challenging task. This paper presents an integrated system that uses YOLOv8 for object detection and BLIP (Bootstrapped Language Image Pretraining) for caption generation. The proposed system also incorporates multilingual translation and text-to-speech (TTS) modules to improve accessibility and user interaction. The system allows users to upload images and receive real-time captions along with translated text and audio output. This approach enhances usability for visually impaired users and improves human-computer interaction. The results demonstrate that the proposed system provides accurate and efficient image captioning compared to traditional approaches.

Keywords: Image Captioning, Deep Learning, YOLOv8, BLIP, NLP, Computer Vision, Object Detection, Text-to-Speech.

I. INTRODUCTION

Image caption generation is an important research area in Artificial Intelligence that focuses on generating meaningful descriptions for images. With the increasing use of smartphones,

social media platforms, and digital systems, a large amount of visual data is generated daily. Understanding and describing this data automatically helps in various applications such as accessibility for visually impaired individuals, content-based image retrieval, surveillance systems, and intelligent human-computer interaction.Traditional methods relied on manual tagging and simple rule-based algorithms, which were inefficient and time-consumingrelationshipswith advancements in Deep Learning, models can now extract high-level features from images and generate natural language descriptions.In this project, we propose a system that integrates object detection and caption generation to provide accurate image descriptions. The system uses YOLOv8 for detecting objects within the image and BLIP for generating captions. Additionally, translation and speech modules are included to enhance accessibility and usability.

II. LITERATURE SURVEY

Earlier approaches for image captioning used template-based and retrieval-based methods. These methods generated captions using predefined sentence structures or by retrieving similar captions from a dataset. However, they lacked flexibility and failed to generate unique and context-aware descriptions.

With the development of Convolutional Neural Networks (CNNs), image feature extraction improved significantly. CNNs are capable of capturing spatial features from images, making them suitable for computer vision tasks. Later, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were used to generate sequential text based on extracted features.

Although RNN-based models improved caption generation, they had limitations in handling long-term dependencies and complex language structures. Recent advancements introduced transformer-based models such as BLIP, which combine vision and language understanding to generate more accurate and context-aware captions.

YOLO (You Only Look Once) models are widely used for real-time object detection due to their high speed and accuracy. The latest version, YOLOv8, provides improved performance and efficiency. Integrating YOLOv8 with BLIP creates a powerful system capable of detecting objects and generating meaningful captions.

A. Technical Feasibility

The system uses Python, TensorFlow/Keras, and pre-trained models like VGG16. These tools are efficient and widely available.

B. Operational Feasibility

The system is user-friendly and allows users to upload images and generate captions easily.

C. Economic Feasibility

The project uses open-source libraries, reducing development cost.

D. Legal Feasibility

The system ensures ethical use of data and respects dataset copyrights.

III. METHODOLOGY

1. Collect dataset (Flickr8K / Flickr30K)
2. Preprocess images and captions
3. Extract features using CNN (VGG16)
4. Tokenize and pad text data
5. Train LSTM model
6. Generate captions
7. Evaluate model performance

IV. PROBLEM STATEMENT

Understanding image content and converting it into meaningful text is a challenging task. Manual captioning is time-consuming and inefficient.The proposed system aims to automatically generate captions for images using deep learning techniques, improving accuracy and efficiency.

V. SYSTEM ARCHITECTURE

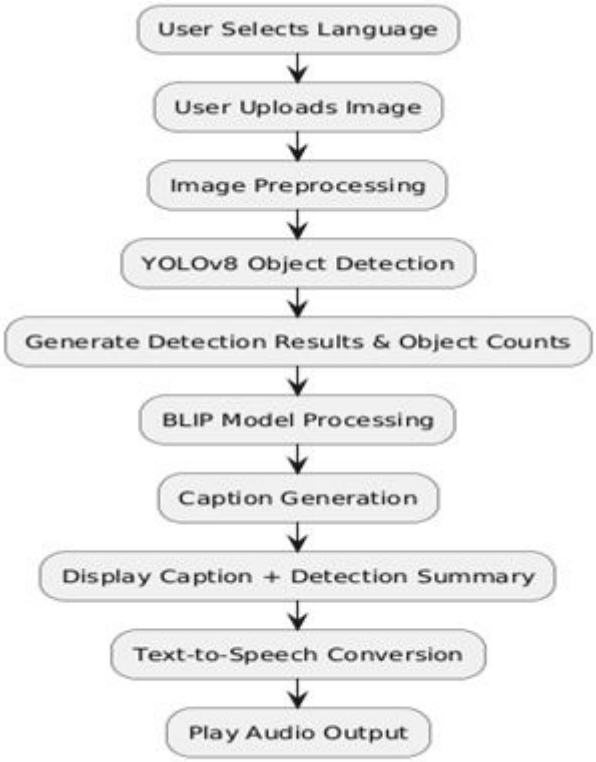
The system consists of three main components:

1.**Image Feature Extraction (CNN)**

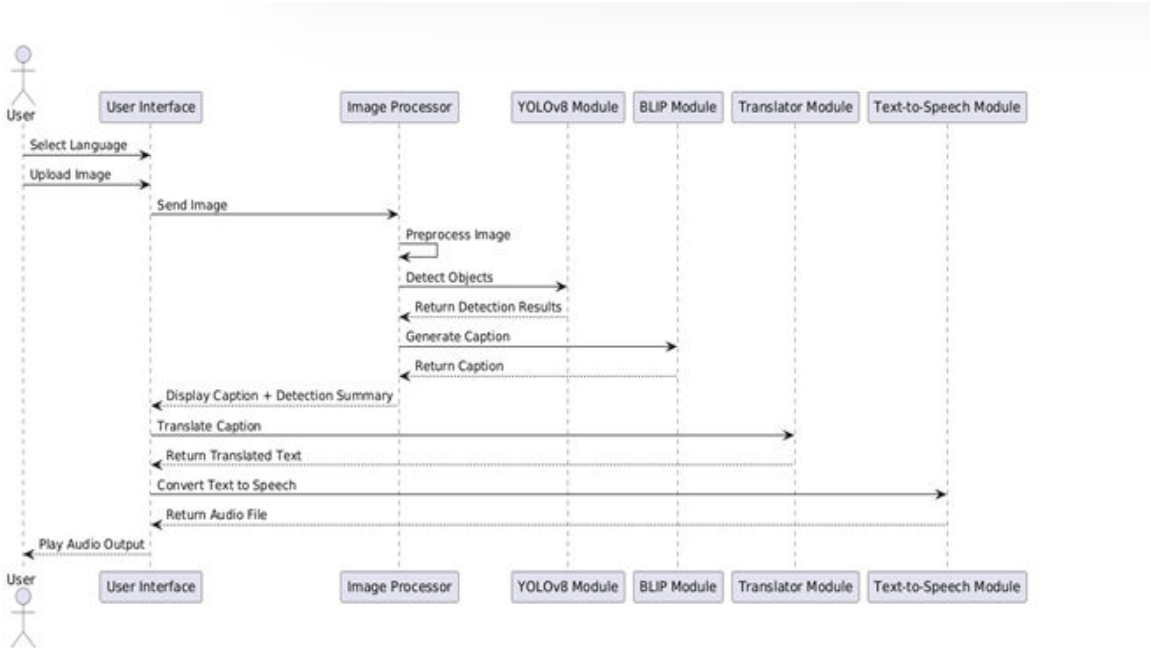
2.**Sequence Model (LSTM)**

3.**Caption GeneratorFlow:**

Image → CNN → Feature Vector → LSTM → C



SEQUENCE DIAGRAM:



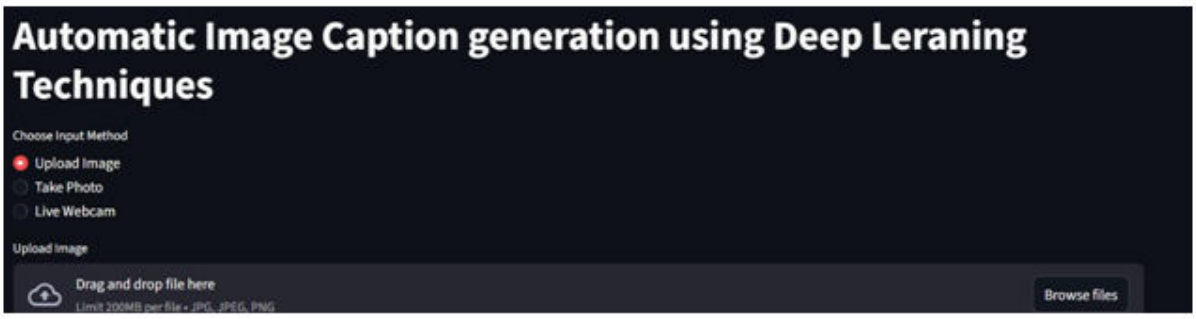
VI. TECHNOLOGY STACK

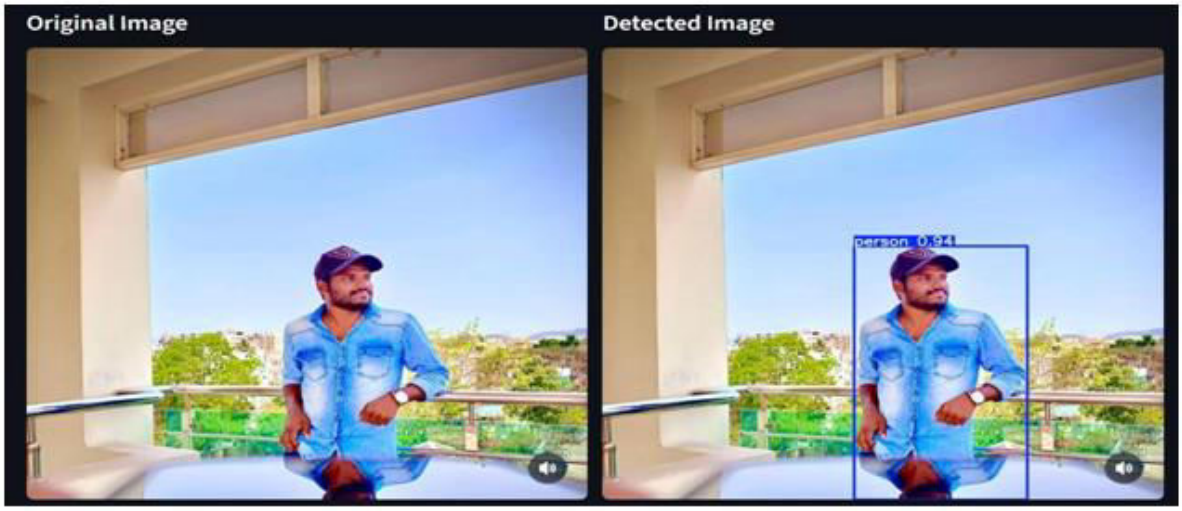
- **Programming Language:** Python
- **Libraries:** TensorFlow, Keras, NumPy, Pandas
- **Model:** CNN (VGG16), LSTM
- **Tools:**VS code/ google colab.

VII. INPUT DESIGN

Input is an image uploaded by the user.

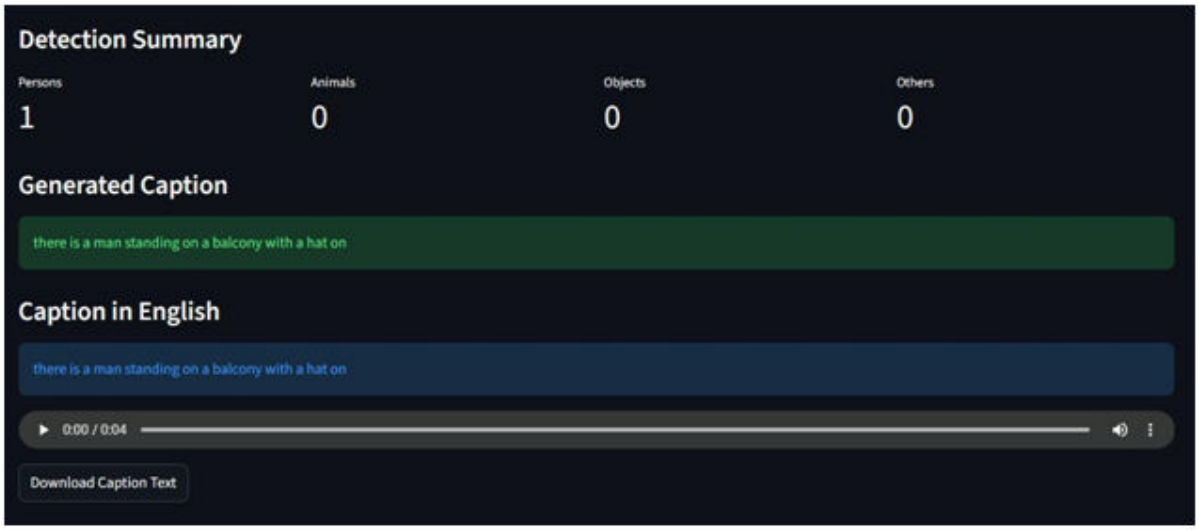
The image is processed and converted into feature vectors using CNN.





VIII. OUTPUT DESIGN

Output is a generated caption describing the image.



Example:

Input → Image of a man

Output → *"there is a man standing on a balcony with a hat on."*

IX. FUTURE SCOPE

- Use Transformer models (like BERT, GPT)
- Improve caption accuracy with attention models
- Real-time captioning for videos

- Integration with mobile applications

X. CONCLUSION

The Automatic Image Caption Generation system successfully combines Computer Vision and NLP to generate meaningful captions.The use of CNN and LSTM models improves accuracy and efficiency. This system has wide applications in accessibility, automation, and AI-based systems.

XI.REFERENCES

[1] Vinyals, O., et al., *Show and Tell: A Neural Image Caption Generator*, IEEE, 2015 https://ieeexplore.ieee.org/document/7298935

[2] Karpathy, A., Fei-Fei, L., *Deep Visual-Semantic Alignments*, 2015 https://cs.stanford.edu/people/karpathy/deepimagesent/

[3] He, K., et al., *Deep Residual Learning for Image Recognition*, 2016 https://arxiv.org/abs/1512.03385

[4] Simonyan, K., Zisserman, A., *VGGNet*, 2014 https://arxiv.org/abs/1409.1556

[5] Ultralytics, "YOLOv8 Documentation," Available:https://docs.ultralytics.com/

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[7] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[8]PyTorch Team, "PyTorch: An Open Source Machine Learning Framework," Available:https://pytorch.org/

[9] Google Text-to-Speech API (gTTS) -https://gtts.readthedocs.io/