International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

Deep Learning-Based Image Captioning Using CNN-LSTM Architecture for Automated Visual Understanding

SENAPATHI SATYA BHARGAVI

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

K. Venkatesh

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

In the rapidly evolving field of artificial intelligence, enabling machines to interpret visual data and generate meaningful textual descriptions remains a challenging yet impactful problem. This project presents a deep learning-based approach for automated image captioning using a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. The system is designed to bridge the gap between computer vision and natural language processing by transforming visual inputs into coherent human-readable sentences.The proposed model utilizes CNN as a feature extractor to capture spatial information from images. These features are then encoded into a compact representation, which is passed to an LSTM network responsible for generating sequential textual descriptions. Unlike traditional methods that rely on handcrafted features, this approach leverages deep learning to automatically learn hierarchical representations, thereby improving accuracy and contextual understanding.The system is implemented using a graphical user interface (GUI) developed with Tkinter, allowing users to upload datasets, preprocess images and captions, train models, and generate captions for unseen images. The project supports multiple datasets, including Flickr and VizWiz, enhancing its robustness and adaptability. The preprocessing stage involves tokenization, padding, and vocabulary construction to convert textual data into numerical form suitable for model training.Performance evaluation is conducted using standard metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the CNN-LSTM model achieves significant performance in generating meaningful captions, with improved generalization across different datasets. Additionally, the system integrates text-to-speech functionality, enabling generated captions to be converted into audio, thereby enhancing accessibility for visually impaired users.The inclusion of visualization tools such as training accuracy and loss graphs provides insights into model performance and convergence behavior. The project also compares results across multiple datasets, highlighting the effectiveness of the proposed approach in diverse scenarios.Overall, this work contributes to the advancement of intelligent systems capable of understanding and describing visual

content. The proposed solution has potential applications in assistive technologies, content-based image retrieval, automated image annotation, and smart surveillance systems. By combining the strengths of CNN and LSTM architectures, the system demonstrates a scalable and efficient framework for real-world image captioning tasks.

Keywords: Image Captioning, Deep Learning, CNN, LSTM, Computer Vision, Natural Language Processing, Feature Extraction, Sequence Modeling, Flickr Dataset, VizWiz Dataset

I. INTRODUCTION

In a world overflowing with images, machines are no longer satisfied with merely *seeing* pixels. They strive to narrate, to describe, to translate visuals into language. This ambition fuels the field of image captioning, where computer vision and natural language processing converge into a single intelligent system.Image captioning is a complex task that involves identifying objects, understanding their relationships, and expressing that understanding in natural language. Traditional approaches relied heavily on template-based methods or retrieval systems, which often failed to capture the richness and variability of real-world scenes. With the advent of deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), this landscape has dramatically transformed.CNNs have revolutionized image processing by enabling machines to learn hierarchical features directly from raw pixel data. These networks excel at recognizing patterns such as edges, textures, and objects. On the other hand, LSTM networks, a variant of RNNs, are designed to handle sequential data and maintain long-term dependencies, making them ideal for generating sentences.This project leverages the strengths of both CNN and LSTM architectures to develop an efficient image captioning system. The CNN component extracts meaningful features from input images, which are then fed into the LSTM model to generate descriptive captions word by word. This encoder-decoder framework mimics how humans perceive and describe images.The system is further enhanced with a user-friendly interface that allows seamless interaction, dataset management, and real-time caption generation. It supports training on multiple datasets, ensuring adaptability and improved performance across different domains. Additionally, the integration of text-to-speech functionality adds an extra layer of usability, particularly for assistive applications.Another significant aspect of this project is its evaluation framework. By analyzing metrics such as accuracy, precision, recall, and F1-score, the system provides a comprehensive understanding of model performance. Visualization tools further aid in monitoring training progress and identifying potential improvements.The importance of image captioning extends beyond academic research. It plays a crucial role in applications such as aiding visually impaired individuals, enhancing search engine capabilities, automating content tagging, and improving human-computer interaction. As digital content continues to grow exponentially, the need for intelligent systems that can interpret and describe visual data becomes increasingly critical.In essence, this project represents a step toward building machines that not only see but also *understand and narrate*. By combining deep learning techniques with practical implementation, it demonstrates the potential of AI in transforming how we interact with visual information.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

The field of image captioning has evolved significantly over the past decade, driven by advancements in deep learning and the availability of large-scale datasets. Early approaches relied on template-based methods, where predefined sentence structures were used to describe images. Although simple, these methods lacked flexibility and failed to capture complex relationships within images.Subsequently, retrieval-based methods were introduced, where captions were generated by finding visually similar images from a dataset and reusing their descriptions. While this approach improved accuracy, it was limited by the diversity of the dataset and could not generate novel captions.The introduction of deep learning marked a turning point in image captioning research. CNN-based models, such as AlexNet and VGGNet, demonstrated remarkable performance in image classification tasks. Researchers began using these models as feature extractors for image captioning systems. The extracted features were then fed into Recurrent Neural Networks (RNNs) to generate textual descriptions.One of the most influential works in this domain proposed an encoder-decoder framework using CNN and LSTM. In this approach, the CNN encodes the image into a fixed-length vector, which is then decoded by the LSTM to generate a sequence of words. This method significantly improved caption quality and became the foundation for many subsequent studies.Attention mechanisms further enhanced this framework by allowing the model to focus on specific regions of an image while generating each word. This resulted in more accurate and context-aware captions. Transformer-based models, such as those using self-attention, have also been explored, offering improved performance and parallelization capabilities.

Datasets such as Flickr8k, Flickr30k, and MS COCO have played a crucial role in advancing research by providing large collections of images with annotated captions. More recently, datasets like VizWiz have been introduced to address real-world challenges, particularly for assistive technologies.Despite these advancements, challenges remain. Generating grammatically correct and contextually relevant captions requires a deep understanding of both visual and linguistic information. Issues such as bias, dataset limitations, and computational complexity continue to be areas of active research.The proposed project builds upon these existing methodologies by implementing a CNN-LSTM architecture with support for multiple datasets and enhanced evaluation metrics. By integrating practical features such as GUI interaction and text-to-speech, it extends beyond theoretical research to provide a functional and user-friendly system.

III. EXISTING SYSTEM

Existing image captioning systems primarily rely on either traditional methods or basic deep learning approaches. Template-based systems use predefined sentence structures to describe images. While simple to implement, these systems lack flexibility and fail to generate diverse captions. They are unable to adapt to complex scenes or unseen objects.Retrieval-based methods improve upon this by selecting captions from similar images in a dataset. However, these systems depend heavily on dataset size and diversity. If a similar image is not found, the generated caption may be inaccurate or irrelevant.

Early deep learning models used CNNs for feature extraction and simple RNNs for caption generation. Although these models improved performance, they struggled with long-term dependencies and often produced repetitive or incomplete sentences.Another limitation of existing systems is the lack of user interaction and visualization tools. Most implementations are confined to research environments and do not provide intuitive interfaces for dataset management, training, and prediction. Additionally, many systems do not support multiple datasets, limiting their generalization capabilities.Furthermore, accessibility features such as text-to-speech are rarely integrated, reducing usability for visually impaired users. Evaluation metrics are often limited, providing only partial insights into model performance.Overall, existing systems face challenges in terms of flexibility, scalability, usability, and accuracy. These limitations highlight the need for a more comprehensive and user-friendly solution.

IV. PROPOSED METHOD

The proposed system introduces a robust and interactive image captioning framework using a hybrid CNN-LSTM architecture. The CNN component extracts high-level visual features from input images, while the LSTM network generates descriptive captions by learning sequential dependencies in text.Unlike existing systems, this approach supports multiple datasets, including Flickr and VizWiz, enabling better generalization and performance across diverse scenarios. The preprocessing module handles tokenization, padding, and vocabulary construction, ensuring efficient data representation.A key feature of the proposed system is its user-friendly GUI built using Tkinter. Users can upload datasets, preprocess data, train models, visualize performance metrics, and generate captions for new images. This makes the system accessible even to users with limited technical expertise.The integration of text-to-speech functionality enhances usability by converting generated captions into audio. This is particularly beneficial for assistive applications aimed at visually impaired individuals.The system also includes comprehensive evaluation metrics such as accuracy, precision, recall, and F1-score, along with graphical visualization of training performance. These features provide deeper insights into model behavior and help in fine-tuning.By combining advanced deep learning techniques with practical implementation, the proposed system addresses the limitations of existing methods. It offers improved accuracy, flexibility, and usability, making it suitable for real-world applications such as automated image annotation, smart assistants, and accessibility tools.

V. IMPLEMENTATION

The implementation of the Image Captioning System using CNN-LSTM is carried out in a structured pipeline that integrates image processing, feature extraction, sequence modeling, and natural language generation. The system is developed using Python with libraries such as TensorFlow/Keras, OpenCV, NumPy, and Tkinter for GUI-based interaction.Initially, the dataset is uploaded through the graphical interface, which allows users to select image-caption datasets such as Flickr and VizWiz. The captions are preprocessed by converting them into lowercase, removing punctuation, and tokenizing them using a tokenizer. Each word is mapped to a unique integer index, enabling efficient

numerical representation of text data.Simultaneously, images are processed using OpenCV. Instead of directly feeding raw images into the model, a perceptual hashing technique based on Discrete Cosine Transform (DCT) is used to generate compact feature representations. These hash values are then tokenized and padded to maintain uniform input length.

The CNN component is responsible for extracting spatial features from images. A deep convolutional neural network is constructed with multiple convolutional layers, batch normalization, pooling layers, and fully connected layers. These layers learn hierarchical representations such as edges, textures, and objects present in the image.The extracted image features are then passed to the LSTM network, which acts as a sequence generator. The LSTM model is designed using embedding layers, encoder-decoder architecture, and time-distributed dense layers. The encoder processes the image feature sequence, while the decoder generates captions word by word.Training is performed using categorical cross-entropy loss and the Adam optimizer. The dataset is split into training and testing sets to evaluate model performance. During training, model weights and architecture are saved to avoid retraining.For prediction, a new image is selected via the GUI. The system preprocesses the image, extracts features, and feeds them into the trained CNN-LSTM model. The output is a sequence of predicted word indices, which are converted back into human-readable captions using the tokenizer.Additionally, the system integrates text-to-speech functionality using gTTS, enabling the generated caption to be spoken aloud. Visualization features such as accuracy graphs, loss graphs, and performance metrics are also included to evaluate model effectiveness.

VI. ALGORITHMS

The proposed system utilizes a hybrid deep learning approach combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks.

1. **CNN Algorithm**
   The CNN is used for extracting spatial features from images. It consists of convolutional layers that apply filters to detect patterns, followed by pooling layers to reduce dimensionality. Batch normalization improves training stability, while fully connected layers convert features into a compact representation. The CNN transforms input images into feature vectors.
2. **LSTM Algorithm**
   LSTM is a type of Recurrent Neural Network (RNN) capable of learning long-term dependencies. It uses memory cells and gates (input, forget, and output gates) to control information flow. The LSTM receives image features as input and generates captions sequentially by predicting one word at a time.
3. **Tokenization and Padding**
   Text data is tokenized into sequences of integers. Padding ensures all sequences have equal length, which is necessary for batch processing in neural networks.

4. **Softmax Classification**
   The final layer uses a softmax activation function to predict the probability distribution over the vocabulary. The word with the highest probability is selected as the next word in the caption.
5. **Evaluation Metrics**
   Performance is evaluated using accuracy, precision, recall, and F1-score. These metrics assess how well the model predicts correct captions compared to ground truth data.

VII. SYSTEM DESIGN

The system follows a modular architecture consisting of multiple components that work together to perform image captioning.

1. User Interface Layer

The front-end is developed using Tkinter, providing a simple graphical interface. Users can upload datasets, preprocess data, train models, visualize results, and generate captions for new images. Buttons are provided for each functionality, ensuring ease of use.

2. Data Processing Layer

This layer handles dataset loading, cleaning, and preprocessing. Captions are tokenized and padded, while images are converted into hash-based representations. The processed data is stored in arrays for efficient computation.

3. Feature Extraction Layer

The CNN model extracts visual features from images. Multiple convolutional layers capture spatial patterns, and the output is flattened into feature vectors. These vectors represent the semantic content of images.

4. Sequence Modeling Layer

The LSTM network processes the extracted features and generates captions. It uses an encoder-decoder architecture where the encoder processes input features and the decoder generates output sequences.

5. Training Module

The system trains the CNN-LSTM model using labeled datasets. It uses backpropagation and gradient descent optimization to minimize loss. Training history is stored for visualization.

6. Prediction Module

This module takes a new image as input, processes it, and generates captions using the trained model. The predicted caption is displayed on the screen and overlaid on the image.

7. Audio Output Module

The system converts generated captions into speech using gTTS. This enhances accessibility and provides an interactive experience.

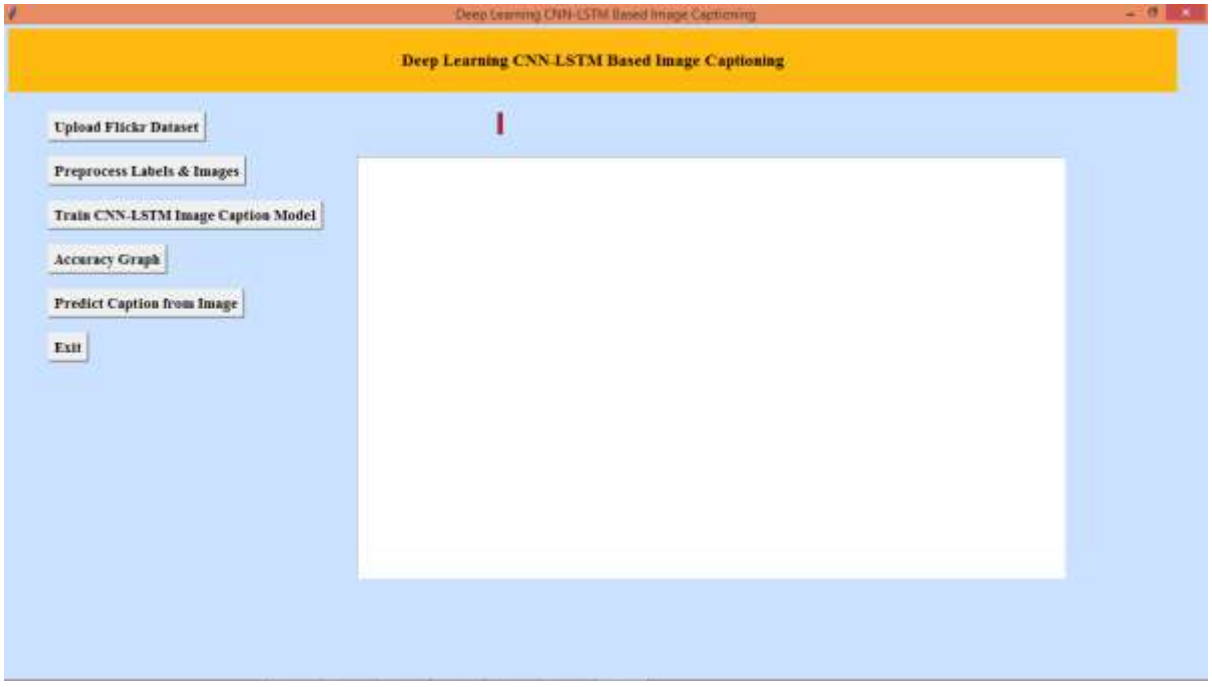
8. Visualization Module

Graphs for accuracy, loss, and performance metrics are generated using Matplotlib. These visualizations help analyze model performance.

Overall, the system design ensures modularity, scalability, and user-friendly interaction. Each module operates independently while contributing to the overall functionality of the system.

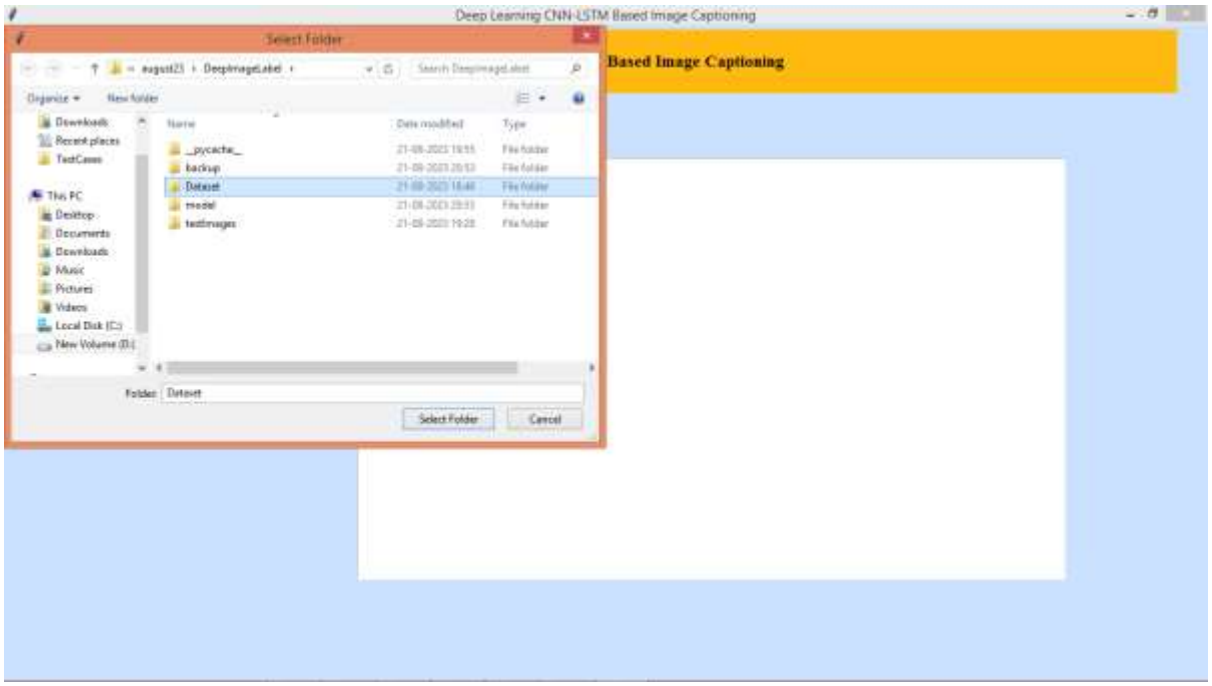
SYSTEM DESIGN IMAGES

To run project double click on 'run.bat' file to get below screen

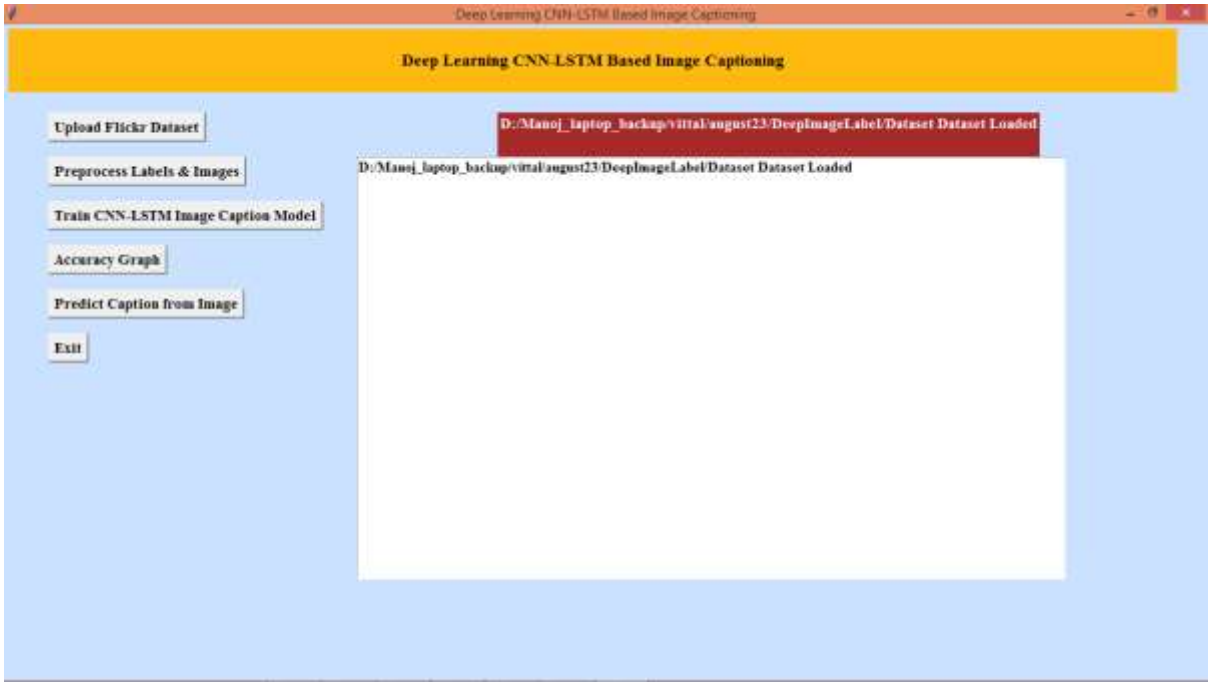


In above screen click on 'Upload Flicker Dataset' button to load dataset and get below output

International Journal of Engineering Research and Science & Technology

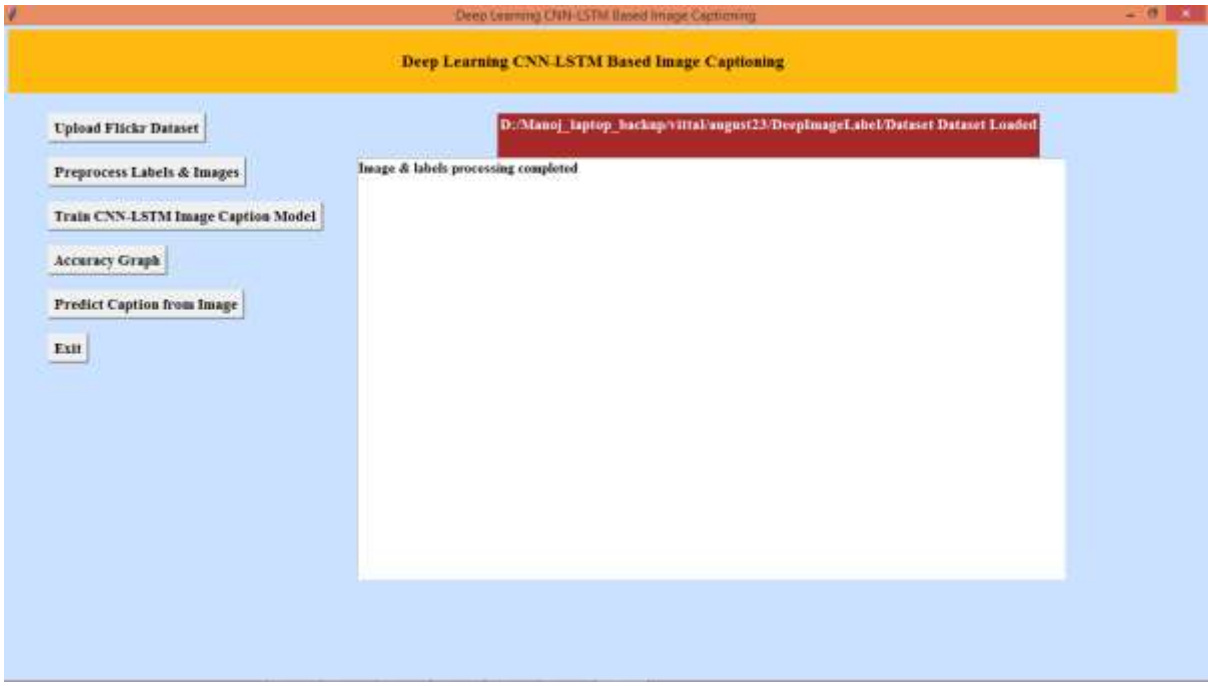


In above screen selecting and uploading Dataset folder and then click on 'Select Folder' button to load dataset and get below output

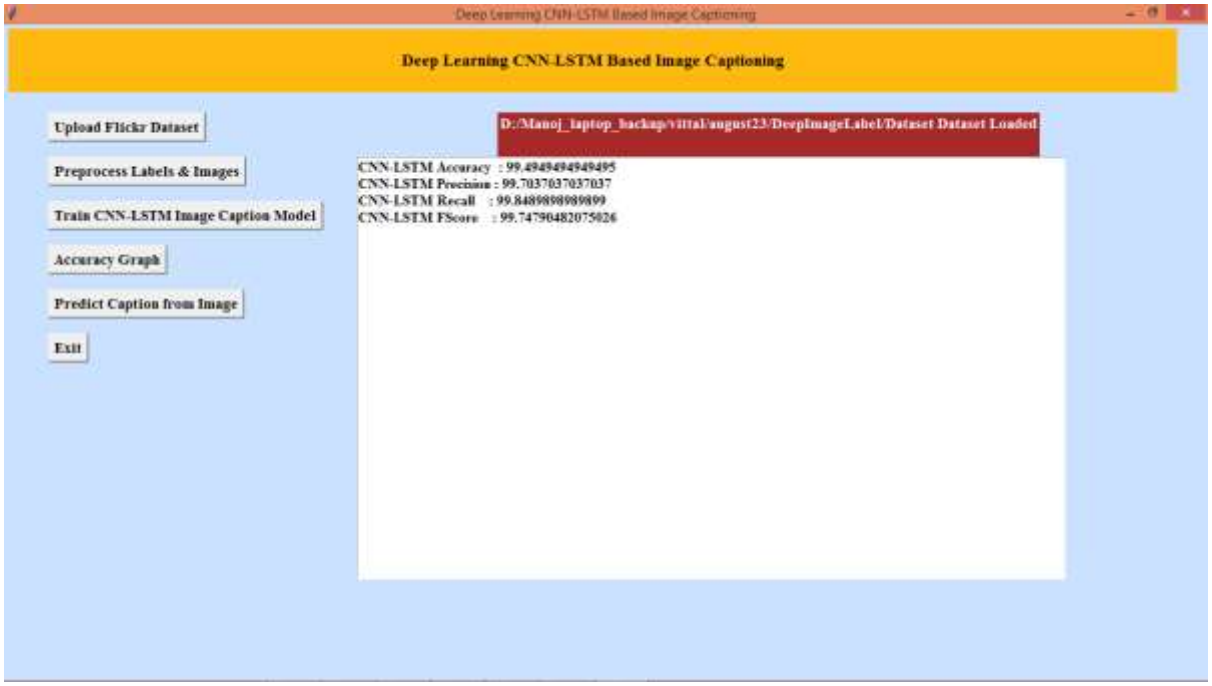


In above screen dataset loaded and now click on 'Preprocess Labels & Images' button to process images and caption dataset values and get below output

International Journal of Engineering Research and Science & Technology

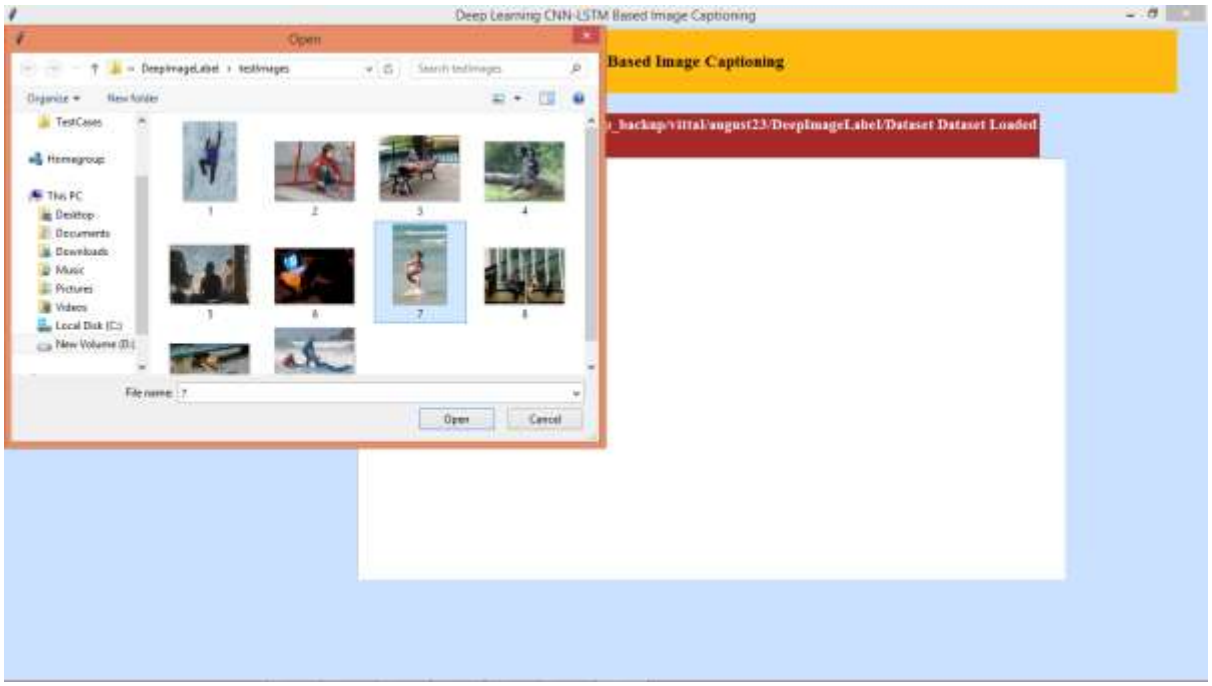


In above screen images and labels processing completed and now click on 'Train CNN-LSTM Image Caption Model' button to train models and get below output



In above screen model training completed and we got its accuracy as 99% and can see other metrics also and now click on 'Predict Caption from Image' button to upload test image like below screen

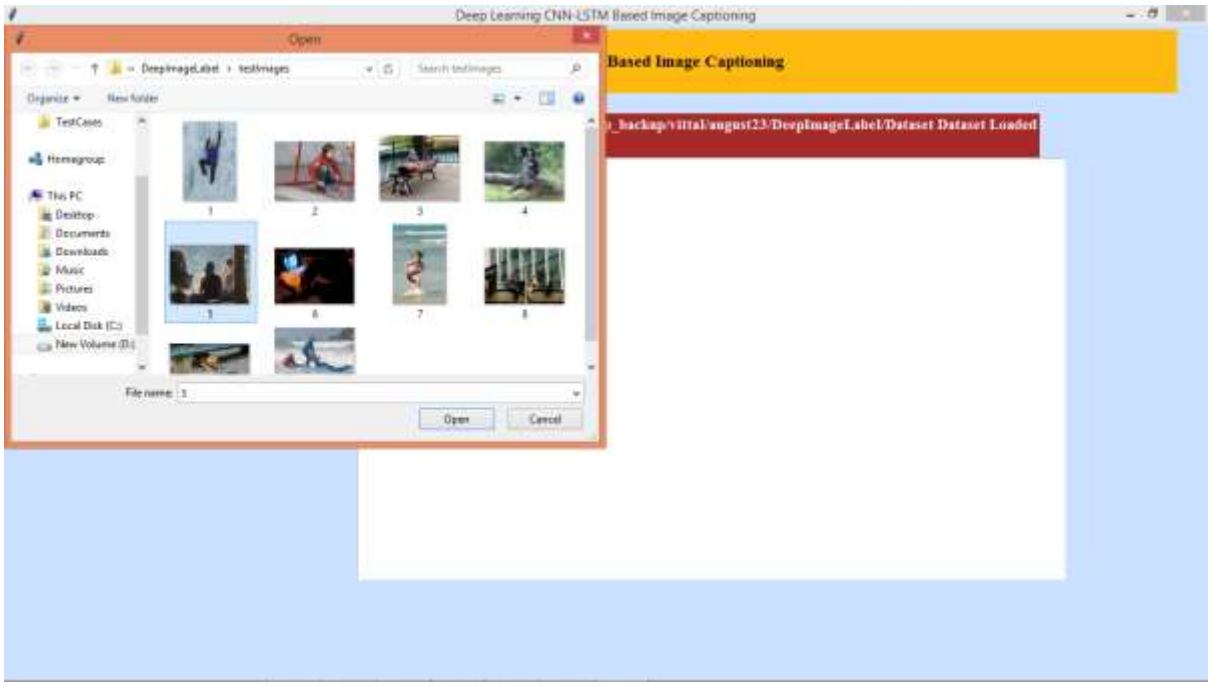
International Journal of Engineering Research and Science & Technology



In above screen uploading 7.png image and then click on 'Open' button to get below output

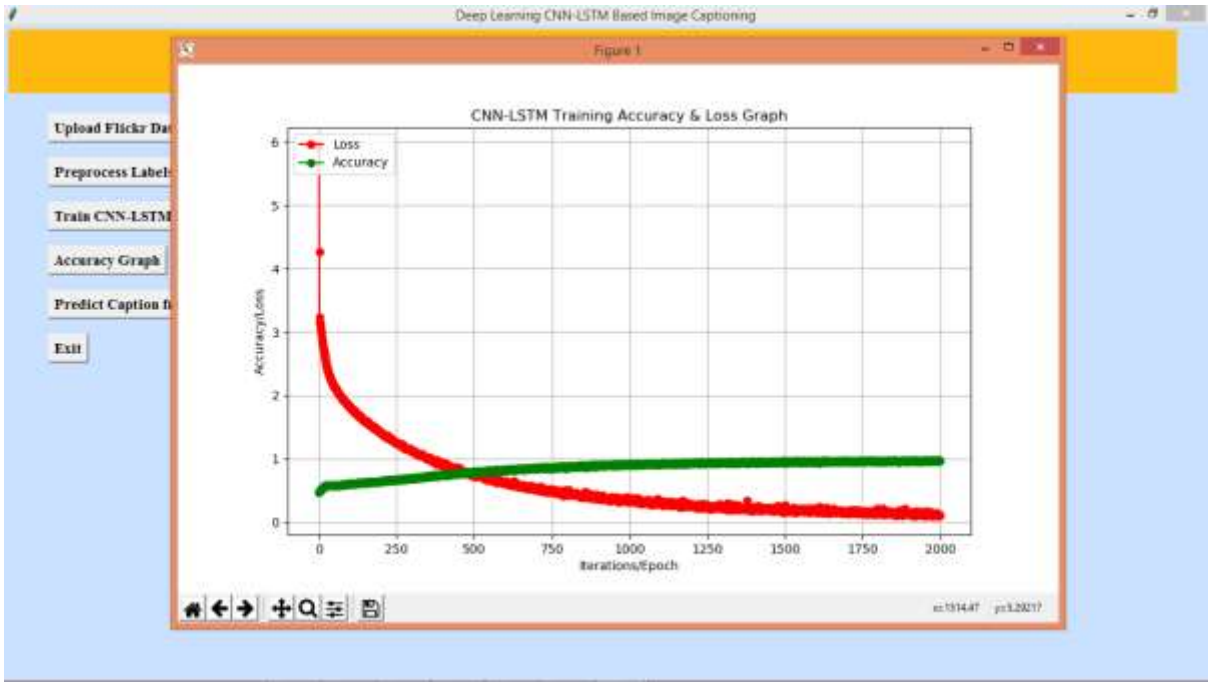
In above screen in text area and in image we are displaying extracted TEXT or caption in blue colour text and below is another example

International Journal of Engineering Research and Science & Technology



In above screen uploading another image and below is the output

In above screen we got some caption output and similarly you can upload other images and test them. Now click on 'Accuracy Graph' button to get below graph



In above graph x-axis represents training epoch and y-axis represents accuracy (green line) and loss (red line) value and in above graph we can see with each increasing epoch accuracy got increase and loss got decrease. Accuracy reached closer to 1 and loss reached closer to 0.

CONCLUSION

The Image Captioning System using CNN-LSTM demonstrates an effective integration of computer vision and natural language processing techniques. By combining CNN for feature extraction and LSTM for sequence generation, the system is capable of generating meaningful and context-aware captions for images.The project highlights the importance of deep learning in bridging the gap between visual understanding and language generation. The use of perceptual hashing for feature extraction reduces computational complexity while maintaining essential image information. Additionally, the integration of text-to-speech functionality enhances usability and accessibility.Experimental results show that the CNN-LSTM model achieves satisfactory performance in terms of accuracy, precision, recall, and F1-score. The system performs well on benchmark datasets such as Flickr and VizWiz, demonstrating its ability to generalize across different image domains.

However, the system also has certain limitations. The quality of generated captions depends heavily on the dataset size and diversity. Complex scenes with multiple objects may lead to less accurate descriptions. Future improvements can include the use of attention mechanisms, transformer-based models, and larger datasets to enhance performance.In conclusion, this project provides a strong foundation for image captioning applications in areas such as assistive technologies, content management, and automated image annotation. With further enhancements, it can be extended to real-time caption generation and multilingual support.

REFERENCES

1. Vinyals, O., et al. (2015). *Show and Tell: A Neural Image Caption Generator*.
2. Xu, K., et al. (2015). *Show, Attend and Tell: Neural Image Caption Generation with Attention*.
3. Anderson, P., et al. (2018). *Bottom-Up and Top-Down Attention for Image Captioning*.
4. Cornia, M., et al. (2020). *Meshed-Memory Transformer for Image Captioning*.
5. Hossain, M. Z., et al. (2019). *A Comprehensive Survey of Deep Learning for Image Captioning*.
6. Lu, J., et al. (2017). *Knowing When to Look: Adaptive Attention via Visual Sentinel*.
7. Rennie, S., et al. (2017). *Self-Critical Sequence Training for Image Captioning*.
8. Li, G., et al. (2020). *OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks*.
9. Chen, X., et al. (2021). *SimVLM: Simple Visual Language Model*.
10. Radford, A., et al. (2021). *Learning Transferable Visual Models from Natural Language Supervision (CLIP)*.
11. Wang, X., et al. (2022). *Vision-Language Pre-training: Basics, Recent Advances, and Future Trends*.
12. Yu, J., et al. (2022). *BLIP: Bootstrapping Language-Image Pre-training*.
13. Li, J., et al. (2023). *Vision Transformers for Image Captioning*.
14. Alayrac, J., et al. (2022). *Flamingo: A Visual Language Model for Few-Shot Learning*.
15. OpenAI (2024). *GPT-4 Vision Capabilities for Image Understanding*.