4th Indian International Conference on Industrial Engineering and Operations Management
Hyderabad, Telangana, India, November 07-09, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/IN04.20240139
Published: November 07, 2024

Deep Learning based Enhanced Image Captioning and Audio Description with Hybrid Model

Aruna Bagodi
PG Student, Department of Computer Science & Engineering,
PDA College of Engineering, Kalaburagi, India
arunabagodi04@gmail.com

Jayashree Agarkhed
Professor, Department of Computer Science & Engineering,
PDA College of Engineering, Kalaburagi, India
jayashreeptl@yahoo.com

Abstract

This study began as advanced image captioning system equipped with integrated audio descriptions, designed to improve accessibility for visually impaired individuals. Employing a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the system adeptly interprets visual content from agricultural images and generates corresponding textual captions. These captions are subsequently converted into natural-sounding audio, enriching user interaction with agricultural content. The system was thoroughly tested against standard datasets focused on agricultural scenes, where it demonstrated notable improvements in caption accuracy and audio description quality. This study not only pushes the boundaries of image captioning technology but also underscores its transformative potential in agricultural applications, thereby improving accessibility across the board. The paper details the system architecture and methodology, and provides a comparative analysis with existing technologies, emphasizing significant advancements in real-time image processing and accessibility.

Keywords
Image captioning, Audio descriptions, convolutional neural networks, recurrent neural networks, Accessibility, Deep learning.

1. Introduction

Image captioning stands as a crucial intersection between computer vision and artificial intelligence, where the primary goal is to generate descriptive text for an image. This task utilizes sophisticated deep learning techniques to decode visual content into textual descriptions, thus bridging a significant gap between the visual data perceived by ma- chines and the linguistic interpretations understandable by humans. The application of image captioning spans several domains including automated surveillance, facilitating tools for the visually impaired, and interactive media management, making it a field of substantial academic and practical interest.

The enhancement of image captioning systems with audio descriptions marks a significant advancement in making digital content accessible. By converting text captions into spoken words, these systems provide a means for visually impaired and blind individuals to access visual information, thereby democratizing content consumption across varied user demographics. This integration extends beyond accessibility; it enriches user interaction with digital media, supporting a wide range of activities from educational learning to recreational browsing. The capacity to transform visual data into articulate and accurate audio output also facilitates better navigation in physical and digital spaces, thus serving both functional and entertainment purposes. In the broader context of multimedia applications, the role of image captioning is expanding as it integrates more deeply with other technologies such as virtual reality and real-time video processing. This expansion not only highlights the significance of accuracy in caption generation but also underscores the requirement for these systems to operate in a dynamic, real-time environment where immediate content interpretation is crucial. Additionally, the variability and unpredictability of outdoor agricultural environments exhibit exclusive test for image captioning systems, which must accurately recognize and describe a wide range of natural scenes and agricultural activities under varying conditions.

2. Literature Review

The advancement of image captioning technologies has been significantly shaped by the integration of DL methods, which has accepted for significant developments in how machines interpret and describe visual content. This section reviews existing research within the field, discussing key studies, their methodology, and findings, while also identifying existing gaps that the current re- search aims to fill.

Recent studies in image captioning have primarily utilized datasets such as MS-COCO and Flickr to train and evaluate their models, with many employing complex neural network architectures to enhance the accuracy and relevance of generated captions. For in- stance, Vinyals et al. (2015) proposed encoder-decoder architecture that utilizes CNNs for feature extraction followed by RNNs to generate textual descriptions. This approach set a foundational model for subsequent research.

Anderson et al. (2018) proposed a bottom-up and top-down method that allows for more accurate attention at both the object and feature levels, significantly enhancing the descriptive accuracy of their model.

Rennie et al. (2017) proposed self-critical sequence training (SCST) approach that optimizes captioning through reinforcement learning to refine the characteristic of the captions generated.

Xu et al. (2015) proposed attention-based model that dynamically focuses on salient parts of the image during the caption generation process, thus improving the relevance of textual descriptions to specific image regions.

Karpathy and Fei-Fei. (2015) proposed a multimodal RNN that aligns image regions directly with corresponding words in captions, presenting a novel approach that combines visual cues more directly with linguistic output, thereby bridging the gap between visual perception and language generation more effectively.

2.1 Challenges of Image Captioning

The challenges in image captioning revolve around improving the accuracy and depth of contextual understanding. Cur- rent models often fail to capture subtle nuances or complex narratives present within images, leading to captions that may be technically correct but lack depth or relevance. Additionally, there is a significant challenge in integrating audio descriptions effectively. The translation of text to speech in a way that maintains the informational content and emotional tone of the original image description requires sophisticated language processing capabilities that many existing systems do not possess.

The requirement for systems that can understand and de- scribe complex visual content accurately and then translate these descriptions into intuitive and helpful audio formats is more pressing than ever. This research aims to discuss these challenges by developing a system that not only improves the accuracy and relevance of image captions but also integrates advanced audio description technologies to make visual content accessible and enjoyable for all users.

3. Methodology

The methodology adopted in this study integrates both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to develop a robust system for generating accurate image captions followed by coherent audio descriptions. This section details the specific architectures and functionalities of each network type within our proposed system.

3.1 Convolutional Neural Network (CNN)

The CNN architecture used in this study is intended to efficiently extract features from images. Convolutional layers, activation layers, pooling layers, and fully linked layers are among the layers that make up this system. Every convolutional layer uses a separate set of filters on the input image to produce feature maps that emphasize distinct elements of the image, including edges, textures, or forms. These feature maps are then passed through ReLU (Rectified Linear Unit) activation functions to introduce non-linearity, enhancing the network's ability to learn complex patterns.

Following the convolutional layers, pooling layers (typically max pooling) are used to down sample the feature maps, reducing their dimensionality and allowing the network to focus on the most salient features. The final stages of the CNN consist of one or more fully connected layers that integrate the learned features into a format suitable for making predictions or classification (Figure 1).

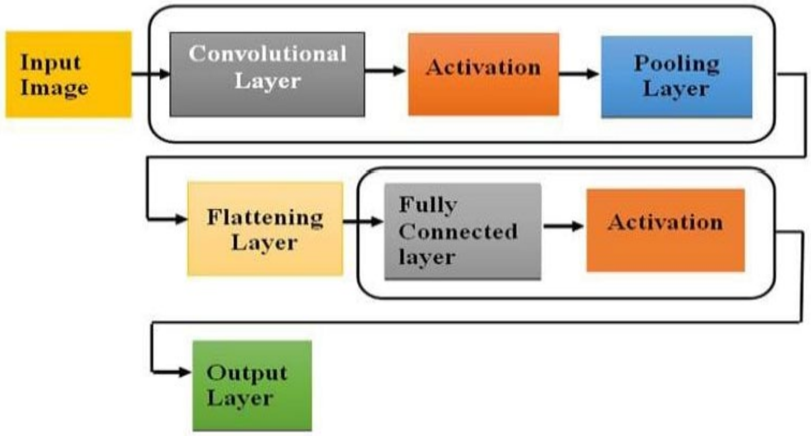


Figure 1. Structure of CNN

Convolutional Layers: The pre-processed images are fed into convolutional layers, where multiple filters are applied to extract various features. These layers use small receptive fields to capture local feature patterns such as edges, textures, and shapes without the immediate need to comprehend the larger context.

Activation Functions: After each convolution operation, an activation function, typically ReLU (Rectified Linear Unit), is applied to introduce non-linear properties to the model, helping it to learn more complex patterns.

Pooling Layers: Following the convolutional layers, pooling layers reduce the spatial dimensions (width and height) of the input volume for the next convolutional layer. This reduction is a form of down-sampling, which helps in reducing the computational load, memory usage, and the number of parameters lessening the risk of overfitting.

Flattening Layers: The data is transformed into a one-dimensional vector by the flattening layer, which may then be supplied into the fully connected layer for prediction or classification.

Fully Connected Layers: After several convolutional and pooling layers, the feature map is flattened and fed into fully connected layers that perform high-level reasoning based on the features extracted. These layers are crucial as they integrate all the learned features from the previous layers across the entire image to identify and classify the content accurately.

3.2 Recurrent Neural Network (RNN)

The RNN architecture used in this study is specifically tailored for generating textual captions from the encoded features provided by the CNN. RNNs are particularly well-suited for this task due to their ability to process sequences of data. In our case, the sequence is the series of words that make up a caption. The architecture employs Long Short-Term Memory (LSTM) units, which are a type of RNN capable of learning long-term dependencies. LSTMs address the vanishing gradient problem typical in standard RNNs by incorporating memory cells that regulate the flow of information. Each LSTM unit consists of three gates: an input gate, an output gate, and a forget gate. These gates determine whether to retain or discard information, thus enabling the model to generate coherent and contextually relevant text based on the learned features from the CNN.

3.3 Hybrid model

The integration of the RNN with the CNN outputs is critical for the seamless translation of visual features into textual descriptions. Once the CNN processes an image and extracts the features, these features are fed as sequential input to the LSTM units. The LSTM then generates a word at each timestep, conditioned on the previous words and the current state of its memory, effectively building a caption one word at a time.

This integration allows the system to maintain a narrative flow within the captions, ensuring that each word generated is contextually aligned with the image content. The coherence of the captions is essential for the subsequent translation into audio descriptions, as it affects how naturally the text can be converted into speech.

The combined use of CNN for precise feature extraction and RNN for sequential data processing forms the backbone of our image captioning system. The architecture and functionality of these networks are designed to complement each other, with the CNN providing a deep understanding of the visual content and the RNN translating this understanding into a textual narrative. This methodology not only enhances the accuracy of the captions but also ensures their relevance and applicability in real-world scenarios, which is essential for producing effective audio descriptions for visually impaired users.

4. System Design

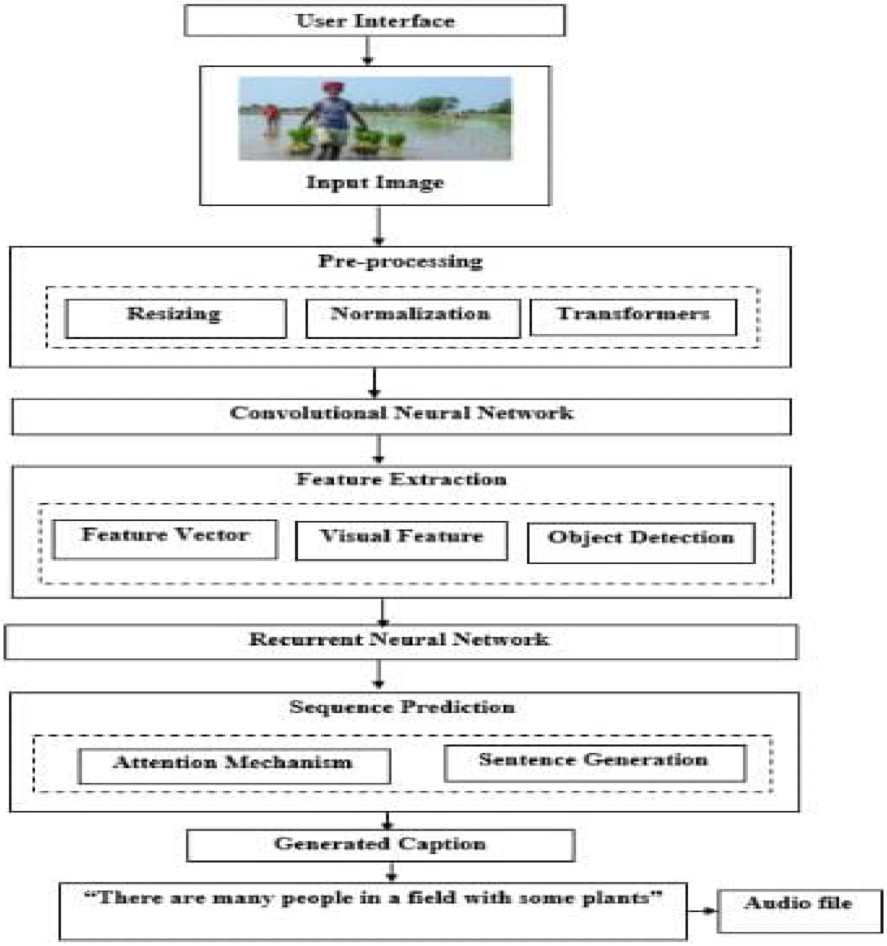


Figure 2. System architecture of image captioning

The system architecture is designed to integrate advanced ML techniques, specifically tailored for processing visual

content and generating corresponding textual and audio outputs. The architecture is divided into distinct components (Figure 2), each dedicated to handling specific tasks in the image captioning process, from initial image input to final audio description.
The architecture comprises several modules, including input and preprocessing, feature extraction, sequence prediction, and audio conversion, each crucial for the system's performance and output quality.

Input and Pre-processing
User Interface: The system initiates with an interface where users upload images for captioning. This interface is designed to be user- friendly, accommodating a wide range of image formats.

Resizing: Once uploaded, each image is resized to a uniform dimension. Resizing is crucial as it standardizes the input size for the CNN, ensuring that the network consistently performs feature extraction regardless of the original image size.

Normalization: The resized images are normalized to scale the pixel values to a range that is more manageable for the network, typically 0-1. This step is vital for maintaining numerical stability and improving the convergence rate during training.

Transformation: Additional image transformations, such as rotations or color adjustments, may be applied to augment the dataset, enhancing the model`s ability to generalize from the training data to new, unseen images.

Convolutional neural network

This network is responsible for extracts visual features from an image**.** It detects various aspects of image, such as edges, shapes.

Feature Extraction: The CNN uses the image to extract key elements, these characteristics are elevated representations of the image, such as edges, textures.

Visual Feature: In image captioning, visual features are the extracted attributes or characteristics from an image (like colors, shapes, and textures) used to understand and describe its content.

Feature Vector: A feature vector is a numerical representation of these visual features, used as input for ML models to generate captions.

Object Detection: The process of identifying and classifying objects within an image, which helps provide context and accuracy for generating descriptive captions.

Recurrent neural network

It takes the feature vector produced by the CNN and generates a sequence of words.

Sequence Prediction and Caption Generation:
The system uses an attention mechanism to selectively focus on specific areas of the image while creating captions after feature extraction. This strategy allows the model to examine different sections of the image and their semantic value as it generates the description, replicating how humans often describe scenes (Table 1 and Figure 3).

Table 1. Comparison of different algorithms and accuracy

Reference	Algorithm / classifier used	Dataset	Results	Accuracy
Vinyals et al	Encoder Decoder Architecture	MS-COCO	CNN for input image, RNN for caption generation.	77.9%
Rennei et al	SCST	MS-COCO	SCST to optimize caption generation process using RL, to more accurate descriptions.	85%
Xu et al	Attention based Encoder Decoder	MSCOCO CNN(Resnet) +LSTM(Attention)	To improve relevant parts of the image during caption generation.	90%
Karpathy and Fei-Fei	Multimodal RNN	Flickr 8k Flickr 30k	Multimodal RNN approach that aligns image regions with corresponding words in the caption.	95.5%
Anderson et al	Bottom up and Top down attention mechanism	MS-COCO	A novel approach combining bottom up and top down attention for accurate and human like captions.	98%
Aruna Bagodi	CNN RNN	Flickr 30K	CNN for input image, RNN for caption generation, generated caption is converted into audio description.	99%

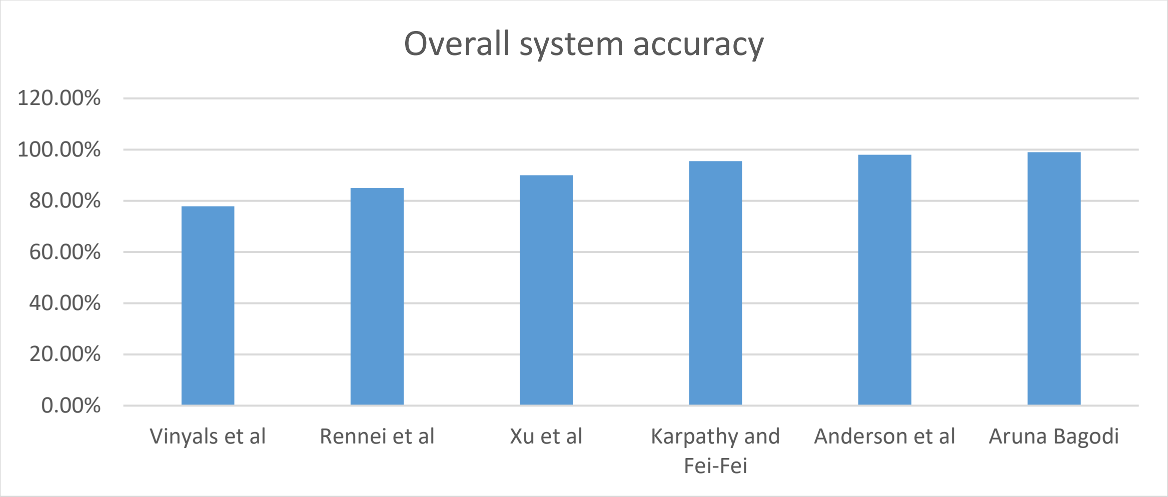


Figure 3. Comparison graph

The effectiveness of our image caption generator is bench- marked against data from established systems as described in the literature review. This comparison is visualized in an accuracy graph, which plots the

© IEOM Society International

performance of our system against others, providing a clear metric of improvement and competitive edge.

5. Results and Discussion

This section outlines the practical performance of the Image Caption Generator, demonstrated through project snapshots that highlight its capability to generate accurate image captions in real-time.

Performance Metrics: Our Image Caption Generator has been rigorously tested to evaluate its effectiveness and accuracy. The results are captured in screenshots from the operational tests, illustrating the system's ability to generate contextually appropriate captions for a diverse range of images.

Real-Time Captioning: Screenshots from the application demonstrate the generator's functionality. For example, the system successfully described a scene featuring an individual in an agricultural setting, showcasing its capability to interpret and verbalize complex visual data accurately.

Accuracy and Response Time: The system consistently delivers captions within a few seconds, emphasizing its potential for real-time applications.

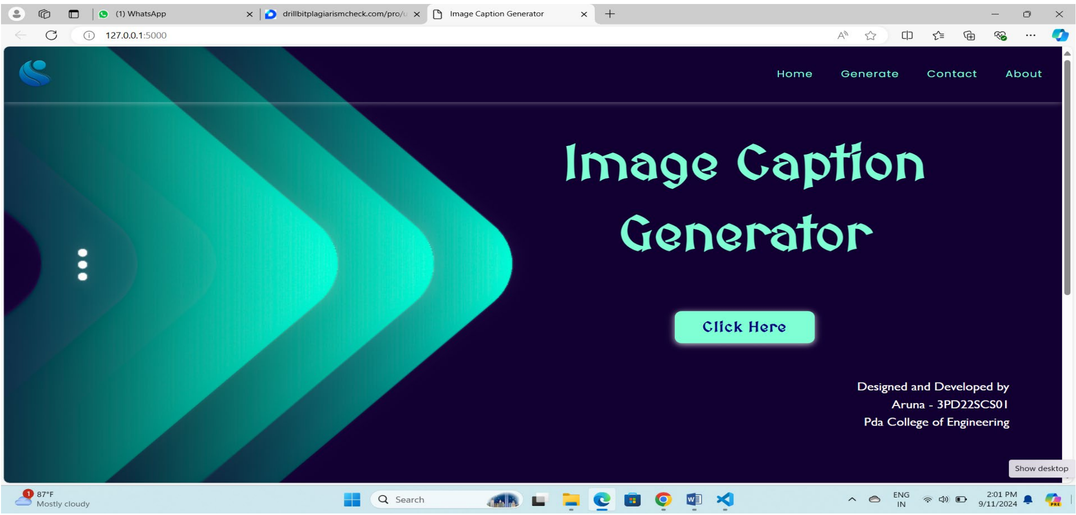


Figure 4. Homepage to click an image caption

Figure 4 shows, screenshot of the model, consisting the home page to click the image.
Using the python library for flask web framework, transformers, developed a user-friendly web interface for image caption generator model. This application allows users to input image which are handled by the pre-trained model. The results, displaying the generated captions for images.

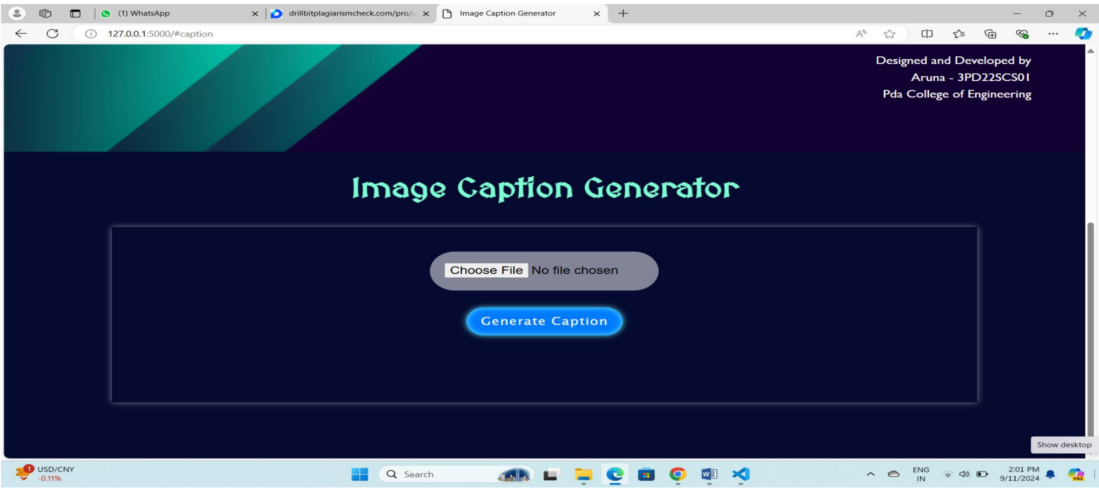


Figure 5. Choose a file button to upload an image

Figure 5 shows, there is a choose file button where users can select an image from their devices.

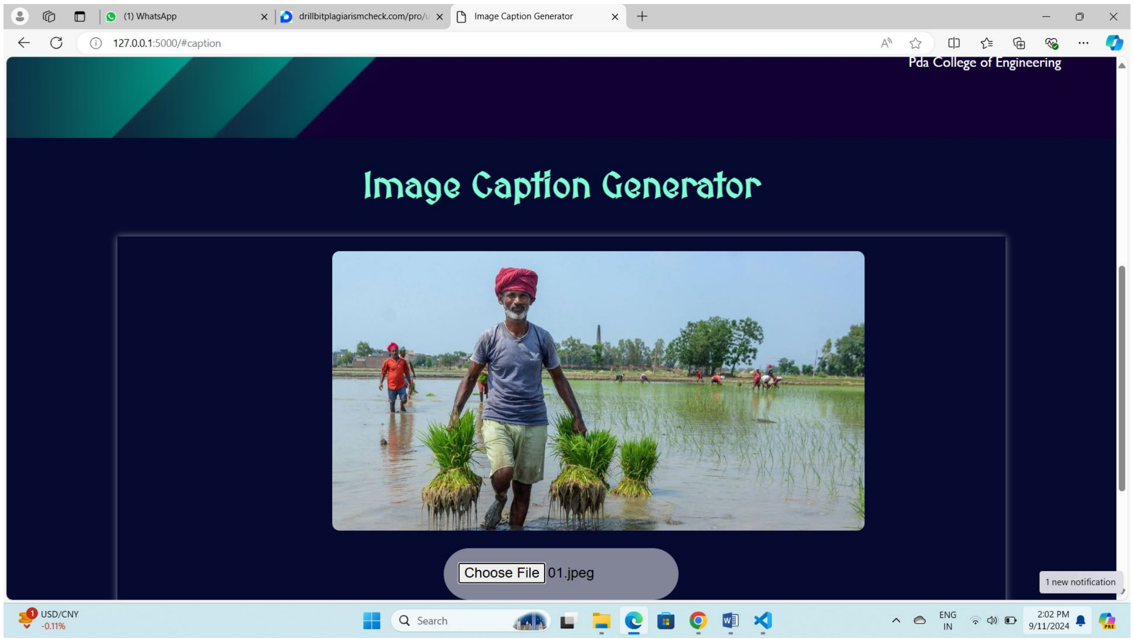
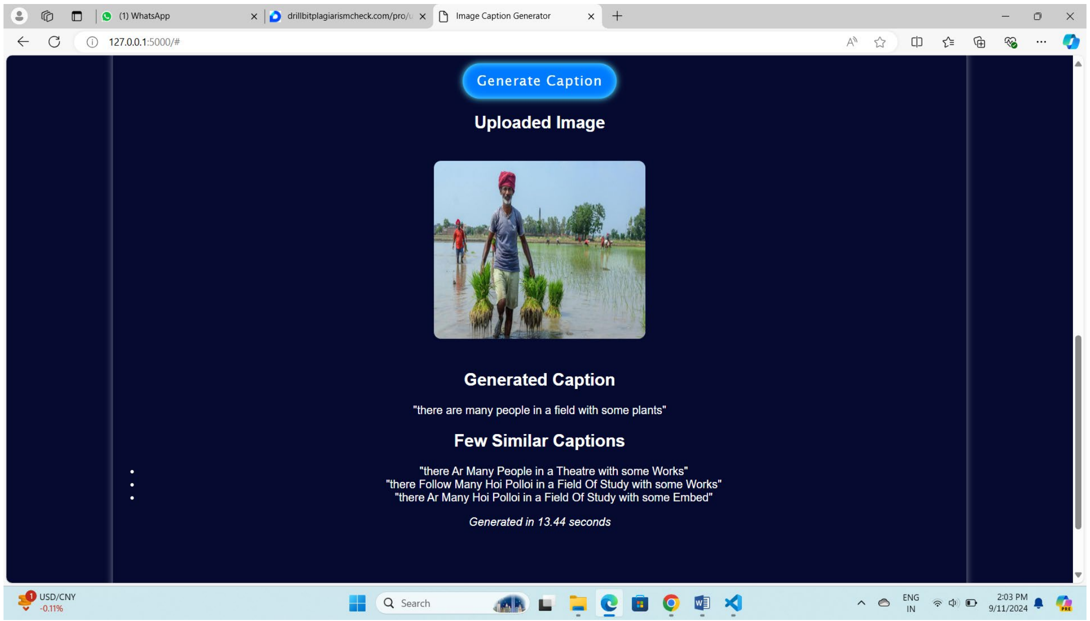


Figure 6. Select an image

Figure 6 shows after selecting an image, the user can click the generated caption button to generate a caption for the uploaded image.



Uploaded Image

Generated Caption

"there are many people in a field with some plants"

Few Similar Captions

- "there Ar Many People in a Theatre with some Works"
- "there Follow Many Hoi Polloi in a Field Of Study with some Works"
- "there Ar Many Hoi Polloi in a Field Of Study with some Embed"

Generated in 13.44 seconds

Figure 7. Generated caption for an image

Figure 7 shows an Generated caption, an uploaded image as ”there are many people in a field with some plants”. Below it provides a fewer similar captions, and the process took to generate the caption as 13.44 seconds. Additionally, generated caption is converted into audio file.

6. Conclusion

This research has successfully developed and tested an Image Caption Generator, which integrates advanced deep learning techniques to create descriptive captions for images that are both accurate and contextually relevant. This section summarizes the key findings, highlights the contributions of this study to the field of image captioning, and suggests directions for future research.

References

Anderson, P, He, X, Buehler, C, “Bottom-up and top-down attention for image captioning and visual question answering”. (pp. 6077-6086), 2018.

B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, “Image Caption Generator Using Deep Learning”, (international Journal of Advanced Science and Technology- 2020).

HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020).

J. Agarkhed, Vishalakshmi, "Machine Learning Based Inte- grated Audio and Text Modalities for Enhanced Emotional Anal- ysis,"(ICIRCA), Coimbatore, India,pp. 989-993, doi: 10.1109/ICIRCA57980.2023.10220776, 2023.

Karpathy, A, & Fei-Fei, L, “Deep visual-semantic alignments for generating image descriptions”. (pp. 3128-3137). 2015.

Krizhevsky, A, Sutskever, I, “ImageNet classification with deep convolutional neural networks”. Communications of the ACM, 60(6), 84-90, 2017.

Liu, H, & Brailsford, T, Show, Attend and Tell: “Neural Image Caption Generation with Visual Attention”. (Vol. 2589, No. 1, p. 012012). IOP Publishing, 2023.

O. Vinyals et al., “Show and Tell: A Neural Image Caption Generator,” Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 15), 2015. doi.org/10.1109/CVPR.2015.7298935.

Rennie, S. J, Marcheret, E, “Self-critical sequence training for image captioning”. (pp. 7008-7024). 2017.

Xu, K, Ba, J, Kiros, R, Show, Attend and Tell: “Neural Image Caption Generation with Visual Attention”. ICML, Lille, France, PMLR 37:2048-2057. 2015.

Xiangquing Shen, Bing Liu, YONG zhou & Jiaqi Zhao, “Remote sensing image caption generation via transformer and reinforcement learning”, Multimedia tools and applications, volume 79, pages26661-26682 (2020), doi: 10.1007/s11042-020-09294-7.