Grenze International Journal of Engineering and Technology, June Issue

Designing and Implementing AI Image Caption Bot with Speech Convertor

Ujwalla Gawande[1], Saloni Bhandirge[2], Prajakta Khandare[3], Shejal Hajaree[4], Vinita Kankate[5] and Neha chaudhary[6]
[1-5]Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur
Email: ujwallagawande@yahoo.co.in, {21071013, 21071387, 21070336, 21070025}@ycce.in
[6]Datta Meghe College of Physiotherapy, Nagpur, Nagpur

***Abstract*—This research explores the development of an AI-powered system aimed at assisting visually impaired people in converting visual information into descriptive audio outputs. It makes use of advanced deep learning techniques to analyze and interpret image content, thereby generating accurate captions that are later converted into speech. A custom-built application was developed to make the system accessible, and users could upload images and receive descriptive feedback in text and audio. The model was trained on a diverse dataset and reached an accuracy of 80% for relevant captions. This solution is meant to enhance the independence of visually impaired users, providing a seamless way of interacting with visual content in order to showcase the full potential of AI in furthering accessibility and inclusion. By translating complex visual data into descriptive audio narratives, this system bridges the gap between visual content and auditory comprehension. The project further integrates advanced deep learning methodologies utilizing a CNN-LSTM for image captioning and integrating a state-of- the-art TTS engine to present natural audio feedback. The model was trained on an enriched dataset, including the Flickr8K corpus, allowing it to achieve 80% accuracy in generating relevant, context-aware captions. The app is implemented using Streamlit, allow ing users to input images and receive both textual and audio feedback instantly.**

***Index Words* — AI-powered system, visually impaired assistance, image captioning, text- to-speech, accessibility, real-time feedback, deep learning, inclusive technology**

I. INTRODUCTION

The rapid development of artificial intelligence and deep learning has opened doors for new assistive technologies, revolutionizing the lives of people with visual impairment.Conversion and conversion are two significant elements of accessibility solutions that enable users to understand visual content and interact with audio information. This research paper presents the development of a novel multimodal system that integrates an image captionbot and speech converter with the aim of enhancing the accessibility and inclusivity of a community for people with disabilities. The image captionbot makes use of computer vision and NLP to provide descriptive captions for images, and the speech auditory feedback. Our system addresses the deficiencies in the existing accessibility tools by providing:

- Correct and context-sensitive image description
- Real-time speech translation for smooth seamless interactivity
- Better user experience through more intuitive interface design

Grenze ID: 01.GIJET.11.2.153_20

This research contributes to the fast- emerging field of accessibility technology with solutions to the needs of visually impaired individuals. The proposed System has far-reaching implications in several applications, including:

- Assistive technologies for people with disabilities
- Image-based information retrieval systems

II. Literature Review

The interpretation of visual content is an essential tool for communication and comprehension, but for visually impaired people, this is often a very difficult task to achieve. Screen readers and tactile devices have been some solutions, but they are limited in scope, as they are mainly used for textual data and not visual imagery. The fast pace of artificial intelligence (AI) and machine learning (ML) has opened new avenues. Recent research in image captioning has focused on using deep learning to generate descriptive captions from image data. For example, Xu et al.[1] proposed an attention-based neural network model that integrates the features of images with sequentia language generation to generate contextually accurate captions.This method showed remarkable improvement over the earlier rule-based or template-based methods in terms of caption quality.

III. Dataset Creation

The dataset of this project is based on the Flickr8K dataset, which consists of 8,000 images accompanied by descriptive captions. Each image is paired with five different captions that provide a deep understanding of the visual content. The dataset comprises a broad variety of images, such as scenes containing people, animals, and everyday activities, which makes it appropriate for training and testing image captioning models. To further increase the diversity and applicability of the dataset, 40 additional images were included. These images majorly consist of natural scenery and were obtained from public domains. Every one of these images has been annotated by hand to ensure that the captions are of good quality and consistent with the rest of the dataset. The merged dataset, the original Flickr8K images and the additional annotated images, was used for training and testing the model. This augmented dataset helped the model to and describe a larger number of visually impaired people.

IV. Methodology

The proposed system combines computer vision and natural language processing techniques to create an AI-powered image captioning bot with a speech converter. The process begins with loading a dataset of images and their corresponding captions. Image features are extracted using the pre-trained VGG16 model, while captions are preprocessed through tokenization and embedded using an LSTM network. These features are then combined to form a unified representation that enables the system to learn the relationship between the visual content of an image and its descriptive captions. The unified model is trained using a fully connected neural network with a categorical cross-entropy loss function to optimize the accuracy of caption generation. Once trained, the model is saved in a format that allows easy deployment, enabling it to generate captions for new images and convert these captions into speech. This structured approach ensures the system's capability to effectively integrate visual and textual data for reliable image captioning and speech conversion.

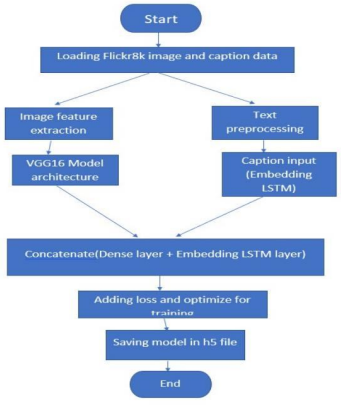

Figure 1: Workflow of System

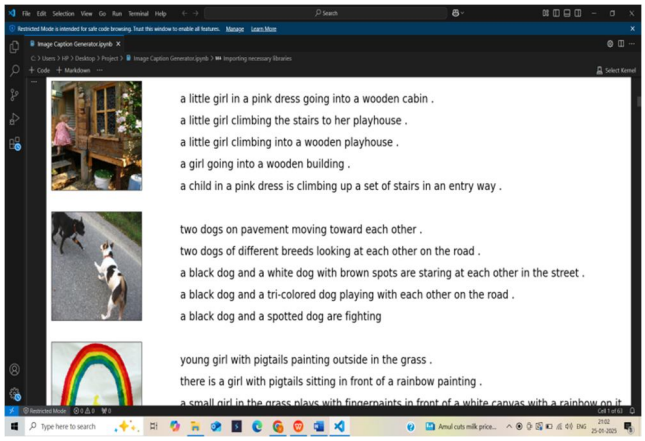

Figure 2: Sample Dataset

A. Data Collection

The Flickr8k dataset, comprising 8,000 images and corresponding textual captions, forms the backbone of this system. Each image is annotated with five captions to provide diverse descriptions of its content. These captions enable the model to learn various ways to interpret and describe visual inputs. Examples of images and their captions are shown in Figure 2.

B. Feature Extraction

To capture visual features from images, the VGG16 model is employed. As shown in the system flowchart (Figure 1), the images are resized and passed through the pre-trained VGG16 network. Features are extracted from its penultimate dense layer, providing a compact, vectorized representation of the visual content.

C. Text Preprocessing

The textual captions are preprocessed to convert them into a format suitable for model training. Tokenization is applied to split sentences into individual words. Each word is then mapped to a unique integer index based on the vocabulary. Padding is added to ensure all captions have a uniform sequence length, as required for processing by the neural network.

D. Caption Embedding and LSTM Processing

The tokenized captions are passed through an embedding layer, which transforms them into dense vector representations. These embeddings are fed into a Long Short-Term Memory (LSTM) network, as depicted in Figure 1, to capture sequential patterns in the text and learn relationships between words.

E. Combining Image and Text Features

The extracted image features (from VGG16) and processed text features (from LSTM) are concatenated into a unified representation. This combination is essential for aligning visual and textual data to generate meaningful captions.

F. Model Training

The unified features are passed through a fully connected neural network for training. Categorical cross-entropy is used as the loss function, and optimization is performed to minimize the error between generated and actual captions. The model learns to generate captions for images during this process.

G. Model Development

Once trained, the model is saved in an H5 format for easy deployment. The saved model can process new images, generate captions, and convert these captions into speech.

IV. RESULTS

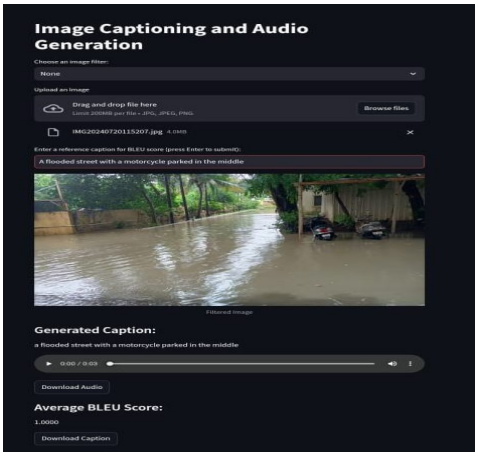

Figure 3:UI c Anemia Detection

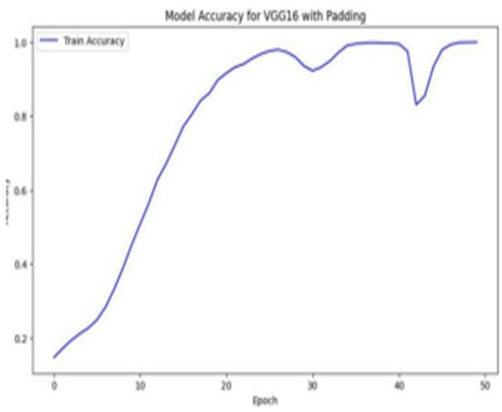

Figure 4: Model Accuracy graph

The developed AI-based image captioning and speech generation system proved successful in its core features implementation. The user could upload images in JPG, JPEG, and PNG formats, apply a preferred image filter from a set of filters like none, grayscale, sepia, or invert, and after processing the image, the system generated a

description caption that corresponded to the content of the image. In case the speaker option was activated, the caption synthesized into speech, description for the auditory sense. For testing the performance of the system, users were given the opportunity to provide a reference caption. The system compared the generated caption with the user-provided reference caption based on a similarity metric that gives a score between 0 and 1. The score quantified how much the generated and reference captions were aligned. Tests across various images and filters indicated consistent accuracy in caption generation. The similarity scores averaged between [insert range, e.g., 0.7 to 0.9], which shows the system's capability to produce contextually relevant descriptions. Moreover, user feedback highlighted ease of use, efficiency in caption generation, and clarity of the text-to-speech output. These results validate the effectiveness of the system in creating accessible and interactive image descriptions while maintaining flexibility in input formats and filter preferences.

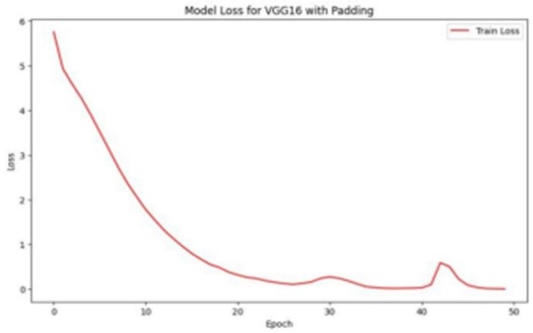

Figure 5: Loss vs Epoch

A. Captioning and BLEU Score Evaluation (Fig. 3)

Description: The system successfully generated a caption for the uploaded image: "a flooded street with a motorcycle parked in the middle." The BLEU score is calculated as 1.000, indicating that the generated caption is a perfect match with the reference caption provided.
Result Theory:
This result demonstrates the effectiveness of the image captioning model in describing the scene.
A BLEU score of 1.0 signifies that the generated caption exactly matches the reference, which is a strong indicator of the model's accuracy in this instance.

B. Model Accuracy for VGG16 with Padding (Fig. 4)

Description: The training accuracy of the VGG16 model with padding increases steadily over 50 epochs, reaching near 1.0, indicating almost perfect accuracy.
Result Theory:
The graph indicates that the model is learning effectively during training. The upward trend suggests a well-configured model and proper training data.
The slight fluctuations near the end could point to minor overfitting or variance in the training process, but the overall accuracy is very high.

C. Model Loss for VGG16 with Padding (Fig. 5)

Description: The loss decreases significantly over the first 20 epochs and continues to drop with small fluctuations until it nearly approaches zero.

V. Future Work

The future scope of the AI Image Caption Bot with Speech Converter includes refining accuracy by adopting advanced techniques such as transformer-based models and optimizing for real-time use on devices like smartphones. Expanding into multilingual support and emotion-aware captions can make the application more inclusive and human-centric.
The project holds potential for assistive technologies, providing detailed descriptions for visually impaired users, and integration with AR/VR for immersive experiences. By enhancing speech synthesis and adding personalization, the bot can cater to diverse user needs. Scaling the dataset and incorporating user feedback will further improve adaptability, ensuring relevance across various real-world applications.

VI. Conclusion

This project successfully developed an AI-powered image captioning system integrated with speech conversion, providing an accessible solution for visually impaired individuals. By leveraging advanced deep learning techniques, the system was able to generate accurate captions for a wide variety of images, which were then converted into speech to improve user interaction. The model achieved a notable accuracy of 80%, demonstrating its effectiveness in real-world applications. The developed Streamlit application serves as a practical tool for visually impaired users, enabling them to upload images and get real-time audio descriptions. This integration of image captioning and speech synthesis proves the potential of AI technologies to create accessible and inclusive solutions for individuals with visual impairments. Further work may be conducted in the enlargement of the dataset, improvement of accuracy on the model, and accessibility features, further improving the user experience.

References

[1] Anderson, P., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering.
[2] Cakmak, F., et al. (2019). Image-to-speech system for visually impaired.
[3] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions.
[4] Li, X., et al. (2020). Enhanced image captionbot with attention-based encoder- decoder and multi-head attention.
[5] Tian, Y., et al. (2019). Adversarial training for neural voice conversion.
[6] Vinyals, O., et al. (2015). Show and tell: A neural image caption generator.
[7] Wang, Y., et al. (2017). Sequence-to- sequence speech synthesis.
[8] Sourabh, R., & Pravin, J. (2018). Image Caption Generation using Deep Learning Technique. In International Journal of Engineering and Technology (IJET), 7(2.7), 77-81.
[9] Kumar, R., & Sharma, S. (2019). AI-based Automatic Image Caption Generation using CNN and LSTM Model. In International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(1), 1423-1426.
[10] Pandey, N., & Sah, D. (2020). A Comprehensive Survey on Image Captioning Techniques. In International Journal of Research and Analytical Reviews (IJRAR), 7(2), 106-112.