SnapSense AI: An Inclusive Image Captioning and Text-to-Speech System

Mrs. Archana V R[1], Mr. Prajwala Gowda P Patil[2], and Mr. Venkateshwara N[3]
[1,2,3]*Member, Jyothy Institute of Technology*

***Abstract*—SnapSense AI is a multimodal system that combines computer vision and Text-to-Speech (TTS) technology to improve accessibility for visually impaired users. The system automatically generates meaningful captions for images and converts them into spoken output, enabling users to understand visual content through audio. By integrating image recognition and natural language processing, SnapSense AI bridges the gap between visual data and human understanding. The generated captions are processed into speech, making digital content more accessible and user-friendly. This approach enhances human–computer interaction and supports inclusive technology design. The system can be applied in assistive tools, education, and real-time information support, contributing to improved accessibility and smarter interaction with visual media.**

***Index Terms*— Accessibility, Computer Vision, Deep Learning, Image Captioning, Natural Language Processing, Text-to-Speech (TTS), Visual Assistance, Visually Impaired Users**

I. INTRODUCTION

SnapSense AI is designed to bridge the gap between visual information and auditory accessibility by addressing the challenges faced by visually impaired individuals in understanding image-based content. In today's digital world, where images are a dominant form of communication, the lack of effective accessibility tools creates a major barrier to information access and independent interaction for users with visual impairments.

Existing solutions such as basic alt text or screen reader support often fail to provide complete and meaningful descriptions of visual content. In many cases, images are labeled with minimal or generic text that does not explain the actual scene, objects, relationships, or context. Additionally, broader accessibility issues such as poor webpage structure, lack of proper labeling, and limited keyboard navigation further reduce usability for visually impaired users. As a result, users often struggle to fully understand or interact with digital content.

SnapSense AI aims to overcome these limitations by developing an intelligent image captioning system integrated with Text-to-Speech (TTS) technology. The system uses deep learning models to generate accurate and context-aware descriptions of images, which are then converted into natural spoken language. This allows users to understand visual content through audio output, improving accessibility and independence.

By generating meaningful and detailed captions instead of simple labels, SnapSense AI enhances the way visual content is interpreted and delivered. It shifts accessibility from passive text reading to active content generation, enabling a more inclusive and user-friendly digital experience. This approach not only improves assistive technology for visually impaired users but also supports broader applications in education, communication, and digital interaction.

II. RELATED WORK

The field of multimodal artificial intelligence has witnessed significant progress in recent years, particularly in the domains of image captioning and Text-to-Speech (TTS) systems. Existing research has primarily focused on improving visual understanding, enhancing language generation, and developing more natural speech synthesis techniques. This section presents a synthesis of related work, highlighting key contributions and their relevance to the SnapSense AI system.

A large body of research in image captioning has explored the integration of computer vision and natural language processing using encoder–decoder

architectures. Early approaches predominantly relied on Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, for sentence generation. These models demonstrated the ability to generate basic image descriptions by learning mappings between visual features and textual sequences. However, they often struggled with contextual understanding, resulting in generic or incomplete captions that failed to capture relationships between objects or the overall scene context.

Recent advancements have introduced attention mechanisms and Transformer-based architectures, which significantly improve the quality of generated captions. Attention models allow systems to focus on specific regions of an image while generating each word, enabling more detailed and context-aware descriptions. Transformer models further enhance performance by capturing long-range dependencies and processing visual and textual information in parallel. Despite these improvements, many existing systems still face challenges in accurately representing complex scenes, emotions, and abstract concepts, particularly in real-world scenarios.

In parallel, research in Text-to-Speech technology has evolved from rule-based and concatenative approaches to modern neural network-based systems. Early TTS systems produced robotic and unnatural speech, limiting their usability in accessibility applications. With the introduction of deep learning models such as Tacotron 2 and WaveNet, speech synthesis has become significantly more natural and expressive. These models are capable of generating high-quality, human-like speech with improved intonation and rhythm. However, challenges such as pronunciation errors, handling of domain-specific terminology, and lack of emotional expressiveness still persist in many systems.

Several studies have also attempted to integrate image captioning with speech synthesis to support assistive technologies. These systems typically combine caption generation models with APIs such as Google Text-to-Speech (gTTS) or Amazon Polly to convert text outputs into audio. While such integrations have proven useful for basic accessibility applications, they often lack contextual depth in captions and produce limited user interaction experiences. Additionally, most existing solutions are not designed with full accessibility in mind, often neglecting intuitive interfaces and multimodal interaction capabilities.

Furthermore, research highlights the importance of dataset quality and diversity in improving model performance. Common datasets such as Flickr8K, MS COCO, and Visual Genome have played a key role in training image captioning models. However, limitations such as lack of domain diversity, linguistic bias toward English, and insufficient representation of complex real-world scenarios restrict the generalization ability of these models. Similarly, TTS systems also suffer from limited multilingual support and difficulties in handling specialized vocabulary.

Overall, the related work indicates that while significant progress has been made in image captioning and speech synthesis individually, there remains a clear gap in achieving seamless, context-aware, and real-time multimodal systems for accessibility. Existing approaches often lack deep integration between vision, language, and speech components, which limits their effectiveness in real-world assistive applications. SnapSense AI aims to address these limitations by combining advanced deep learning models, improved contextual understanding, and natural speech synthesis into a unified and accessible system.

III. ARCHITECTURE DESIGN

The SnapSense AI system follows a deep learning-based encoder–decoder architecture designed to generate image captions and convert them into speech. The architecture is composed of multiple interconnected modules that work together to transform visual input into meaningful audio output.

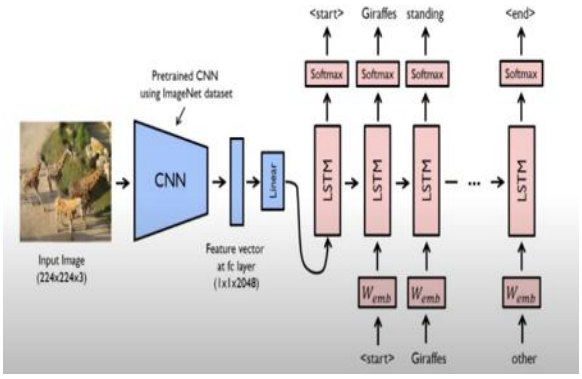


Figure 1: Neural network architecture for image captioning

1.Input Layer: RGB image of size 224 × 224 × 3
Accepts and standardizes the input image for processing. The image is resized and normalized before being passed to the feature extraction stage.

2.Convolutional Neural Network (CNN): Extracts high-level visual features from the input image. A pre-trained CNN model such as ResNet, VGG, or Inception (trained on ImageNet) is used.The CNN processes the image and generates a compact feature representation from its fully connected layer. A deep feature vector representing objects, patterns, and spatial relationships in the image.

3.Linear (Fully Connected) Layer: Transforms CNN output into a suitable representation for sequence generation. Reduces and maps the high-dimensional CNN feature vector into a lower-dimensional embedding space. A feature vector (e.g., 1 × 2048) passed to the language model.

4. Word Embedding Layer: Converts input words into dense vector representations. Each word in the vocabulary is mapped to a fixed-size embedding vector that captures semantic meaning. These embeddings help the model understand relationships between words in a continuous vector space.

5. LSTM (Long Short-Term Memory Network) – Sequence Generation: Generates the image caption word by word. The LSTM acts as the decoder in the architecture.
It takes two inputs: Image feature vector (from CNN + Linear layer), Previous work embeddings. The generation process starts with a <start> token and continues sequentially until an <end> token is produced. Stacked LSTM layers improve learning of long-term dependencies in sentence construction.

6. Softmax Layer: Predicts the next word in the sequence. Computes probability distribution over the entire vocabulary. The word with the highest probability is selected as the next word in the caption generation process. This process repeats until the final caption is generated.

IV. METHODOLGY

The Snap Sense AI system is designed to help visually impaired users understand image content through automatically generated captions and speech output. The system follows a sequential workflow consisting of image processing, caption generation, and text-to-speech conversion.
Initially, the user uploads an image through the application interface. The uploaded image undergoes preprocessing operations such as resizing, normalization, and quality enhancement to ensure consistency for further processing. A pre-trained Convolutional Neural Network (CNN), such as ResNet, is then used to extract meaningful visual features from the image.
The extracted features are passed to a caption generation model based on Long Short-Term Memory (LSTM) networks or Transformer architecture. This model analyzes the visual information and generates a meaningful textual description that accurately represents the objects, actions, and context present in the image.
Once the caption is generated, it is forwarded to the Text-to-Speech (TTS) module. Advanced speech synthesis techniques convert the textual description into clear and natural-sounding audio. The generated voice output is then played through speakers or headphones, enabling visually impaired users to understand the image content without relying on visual information.
The overall methodology ensures a smooth transition from image understanding to audio-based communication, thereby improving accessibility and user experience.

V. RESULT ANALYSIS

Figure 2: Flower as Input Image

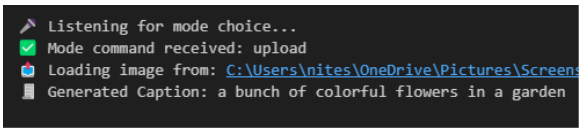
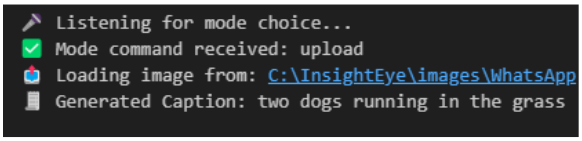


Figure: Input Image2



The proposed SnapSense AI system was evaluated based on its ability to generate meaningful image captions and convert them into understandable speech output. The system successfully identified major objects and activities present in test images and generated descriptive captions that conveyed the overall context of the scene.

The CNN-based feature extraction module effectively captured important visual details, while the caption generation model produced grammatically correct and contextually relevant descriptions. The integration of the Text-to-Speech module enabled these descriptions to be delivered as natural audio, making the information accessible to visually impaired users.

Experimental observations indicated that the system performed well for common real-world images containing people, animals, vehicles, and everyday objects. The generated speech was clear and easy to understand, contributing to a positive user experience. However, a few limitations were observed. Complex images containing multiple objects, unusual scenes, or domain-specific content occasionally resulted in less detailed captions. The accuracy of the output also depended on image quality and the diversity of the training dataset. Despite these challenges, the system demonstrated promising performance as an assistive technology solution.

VI. CONCLUSION

SnapSense AI presents an effective and user-friendly solution for improving accessibility to visual content. By combining image captioning techniques with text-to-speech technology, the system enables visually impaired individuals to receive meaningful descriptions of images in an audio format.

The project successfully integrates computer vision, natural language processing, and speech synthesis into a unified framework. The generated captions provide contextual information about images, while the speech output ensures easy access to that information through auditory means.

The proposed system contributes to inclusive technology development by helping bridge the gap between visual information and auditory understanding. Future enhancements may include multilingual support, real-time object detection, emotion recognition, and improved caption generation using advanced Transformer-based models to further increase accuracy and usability.

REFERENCES

[1] X. Wang, W. Li, C. Chen, and J. Wang, “Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7349–7361, 2020, doi: 10.1109/TGRS.2020.2965234.

[2] Y. Zhang, H. Wang, and X. Lu, “A Joint-Training Two-Stage Method for Remote Sensing Image Captioning,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023, doi: 10.1109/LGRS.2022.3225561.

[3] Hossain, M. A. Hossain, and F. Muhammad, “Automatic Image and Video Caption Generation with Deep Learning: A Concise Review and Algorithmic Overlap,” *IEEE Access*, vol. 8, pp. 218386–218407, 2020, doi: 10.1109/ACCESS.2020.3042281.

[4] G. M. Snoek and M. Worring, “Modelling Semantic Aspects for Cross-Media Image Indexing,” *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1460–1473, Oct. 2007, doi: 10.1109/TMM.2007.906579.

[5] Z. Zhang, W. Diao, X. Sun, Y. Yan, and K. Fu, “NWPU-Captions Dataset and MLCA-Net for

Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, doi: 10.1109/TGRS.2022.3195298.

[6] L. Huang, W. Wang, J. Chen, and X. Wei, "Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation," *IEEE Trans. Multimedia*, vol. 23, pp. 3174–3184, 2021, doi: 10.1109/TMM.2020.3018973.

[7] Y. Li, X. Lu, and W. Diao, "VLCA: Vision-Language Aligning Model With Cross-Modal Attention for Bilingual Remote Sensing Image Captioning," *Big Data Mining and Analytics*, vol. 6, no. 1, pp. 77–89, Feb. 2023, doi: 10.26599/BDMA.2022.9020021.

[8] J. Wang, Y. Zhang, and Z. Wang, "Cross-Lingual Image Caption Generation Based on Visual Attention Model," *IEEE Access*, vol. 8, pp. 114764–114773, 2020, doi: 10.1109/ACCESS.2020.3003560.

[9] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3321–3334, 2020, doi: 10.1109/TGRS.2019.2957251.

[10] W. Diao, X. Sun, F. Dou, M. Yan, and K. Fu, "Multiscale Multi-Interaction Network for Remote Sensing Image Captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3058552.