Journal of Theoretical and Applied Information Technology
31st May 2025. Vol.103. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LEVERAGING SPEECH FOR DYNAMIC IMAGE CAPTIONING: A MOBILENETV3 AND LSTM APPROACH

[1]PREETY SINGH, [2]NAGA DURGA SAILE K, [3]TAKKEDU MALATHI, [4]T RAVI, [5]DIPAK J DAHIGAONKAR, [6]CHUNDURI LAVANYA

[1,2,]VNR Vignana Jyothi Institute of Engineering &Technology, Hyderabad, India
[3]Aurora Deemed to be University, Hyderabad, India
[4] Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology Avadi,Chennai,Tamilnadu, India
[5]Ramdeobaba University, Katol Road, Nagpur, India
[6]KL University, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India
E-mail: [1] preeti17singh@gmail.com, [2]saileknd3@gmail.com, [3]malathi@aurora.edu.in , [4]drravit@veltech.edu.in, [5]dahigaonkardj@rknec.edu, [6]lavanyabbch@kluniversity.in

ABSTRACT

This paper presents a novel way of generating captions for images using an automatic image captioning system; the proposed model combines MobileNetV3 and LSTM to create captions that are accurate and relevant to the image being depicted. MobileNetV3 acts as a feature extractor as it extracts vital components of pictures at a reasonable computational cost. These features are then taken to the LSTM network and from it, descriptive captions from visual context are made. As a measure that improves user access, the generated captions are further translated to sound using Google Text-to-Speech (GTTS), which is especially important for the visually impaired and other hand-free users. Cross-sectional experimental assessments of the performances of the proposed model on the Flickr8k dataset further validate the impressive usefulness of the proposed model for generating faithful and comprehensive descriptions for media items that are potentially useful in assistive technologies, media organizing, and interactive systems.

Keywords: *MobileNetV3, Long Short-Term Memory (LSTM), Google Text-to-Speech (GTTS)*

1. INTRODUCTION

With the growing need for assistive technologies, image captioning systems have become a crucial tool for bridging the gap between visual and textual information. These systems automatically generate descriptive text from images, enabling applications in accessibility, multimedia organization, and human-computer interaction. For visually impaired (VI) individuals, integrating speech synthesis with image captioning can significantly enhance usability, making visual content more accessible. However, existing solutions face critical limitations: (1) Many models prioritize text-based descriptions without incorporating real-time audio output, limiting accessibility. (2) State-of-the-art captioning models, particularly Transformer-based architectures, demand high computational resources, restricting real-world deployment on low-power devices.

To address these gaps, this study proposes a novel image captioning framework that integrates MobileNetV3 and Long Short-Term Memory (LSTM) networks for efficient, lightweight caption generation. MobileNetV3 is a computationally optimized Convolutional Neural Network (CNN) designed for real-time applications, ensuring fast and accurate visual feature extraction. These extracted features are then processed by an LSTM network, which excels in generating coherent and contextually rich captions. Additionally, we incorporate Google Text-to-Speech (GTTS) to convert textual captions into speech, creating a multimodal system that enhances accessibility. This combination makes our model particularly valuable for visually impaired users and hands-free applications.

1.1 Problem Statement & Research Questions

Despite advancements in deep learning-based image captioning, existing models often fall short in terms of accessibility and computational efficiency. Traditional architectures like ResNet and Inception, while powerful, are computationally expensive and unsuitable for real-time deployment on resource-constrained devices. Moreover, while speech integration has been explored in some studies, few have systematically optimized both text and audio generation for real-world accessibility applications.

This research aims to answer the following key questions:

- Can lightweight models like MobileNetV3 provide competitive performance in image captioning compared to complex Transformer-based models?
- Does the addition of GTTS significantly enhance accessibility for visually impaired users in real-time applications?

1.2 Criteria for Literature Screening

To establish the research foundation, conducted a comprehensive literature review spanning studies from 2017 to 2024. Our screening criteria included:

- Use of deep learning models for image captioning, with a focus on CNN-RNN and Transformer-based approaches.
- Datasets used in prior studies, including benchmark datasets such as Flickr8k, MS-COCO datasets.
- Comparison of architectures, novel techniques, or performance evaluation using BLEU, METEOR scores.

1.3 Research Gaps & Unique Contributions

While prior studies have explored image captioning and speech synthesis separately, few have systematically combined an optimized lightweight CNN (MobileNetV3) with LSTM for efficient caption generation and real-time speech conversion. Our work fills this gap by demonstrating that a computationally efficient model can achieve competitive performance without requiring high-end hardware.

Key contributions of our study include:

- Development of a lightweight image captioning system leveraging MobileNetV3 and LSTM, optimizing both speed and accuracy for real-time applications.
- Integration of Google Text-to-Speech (GTTS), transforming textual captions into audio, thereby enhancing accessibility for visually impaired users.
- Identification of literature gaps related to resource-efficient image captioning with speech output, providing new research directions for accessibility-driven AI systems.

Through extensive experimentation on the Flickr8k dataset, we demonstrate that our model generates accurate and contextually meaningful captions while significantly reducing computational overhead. This research contributes to the advancement of assistive technologies, AI-driven accessibility solutions, and multimodal interaction systems

2. RESEARCH METHOD

More recently, deep learning methodologies have been considered critical to drive image captioning systems forward in their ability to handle image information, thus enhancing captions' functionality and accessibility in applications with visual content. There are different techniques used which all aim at improving the quality of captions and time concerns as well as relevance of captions in specific contexts. The following provides an overview of the methods employed in image captioning literature, their comparison, and a discussion of their significance. A Deep Learning approach initiated in image captioning is based on CNN-LSTM encoder-decoder frameworks; these are more fitting for generating descriptions for illustrations with sections for feature regeneration from the images in tandem with sequence prophesying. Despite the effectiveness of CNN-LSTM models, they often struggle with handling complex visual relationships and require improvements in efficiency, particularly for real-time applications and low-power devices. Moreover, existing studies have mostly focused on caption accuracy but have overlooked accessibility enhancements such as integrated speech synthesis.
He et al. [1] review deep learning techniques for image captioning and argue that attention is an interesting achievement as it enriches the primary model by adapting the focus for matching image and text features. Tavakoliy et al. [2] applied a saliency-based attention mechanism that further improves the harmony between assigned attention by the machine and human vision and significantly increases the contextuality of captions earned by

models. Mathur et al. [3] focused on real-time captioning on handheld devices using simplification on the encoder-decoder system so they could process quickly and make precise without needing to affect caption quality. However, these studies mainly emphasize mobile-based optimization and do not fully explore the potential of integrating multimodal accessibility features, such as text-to-speech conversion, which could enhance usability for visually impaired users. These efforts focus on the potential of extending mobile captioning options to enhance the accessibility and usability of mobile devices, especially in finding new ways for assistive technology.

Object detection-based captioning elaborates on the approaches for captioning and increases the accuracy of captioning by detecting and describing multiple objects and attributes of an image. Amritsar et al. [4] used CNN-LSTM architectures to overcome the difficulties in image captioning by improving caption quality and providing better access for visually impaired customers. Kinghorn et al. [5] applied a region-based captioning methodology for IAPR TC-12, which would provide more descriptive information by describing specific segments of an image. These object detection-based approaches help to emphasize the effectiveness of the region-specific description in the captions, more importantly in the description of the specific areas of images, like in medical imaging, self-driving car navigation, and surveillance systems. Despite their improvements in fine-grained captioning, these methods remain computationally intensive and are less suited for real-time applications. Additionally, they lack integration with speech-based assistive technologies, which could further extend their practical applications.

Lu et al. [6] aimed towards the task of remote sensing image captioning while introducing the RSICD dataset to environmental and resource management tasks. Collectively, these writings suggest the improvement the image-captioning technology makes, advancing from a canonical architecture that was CNN-RNN architecture to an enriched Transformer system working with attention. The transition from CNN-RNN models to Transformer-based architectures has significantly improved caption accuracy and contextual understanding. However, Transformer-based models such as Object Relation Transformer (ORT) proposed by Hossain et al. [7] and the bottom-up attention module by Ding et al. [10] require high computational power, limiting their usability in resource-constrained environments. Our study aims to bridge this gap by leveraging a lightweight MobileNetV3-based approach that ensures both efficiency and high-quality captioning performance. Kesavan et al. [8] and Hani et al. [9] selected CNN-RNN models with attention mechanisms for image captioning tasks on MS-COCO since they get a better score than the model that can just pass over small picture details. Ding et al. [10] proposed a bottom-up attention module, which allows the better integration of high and low levels of image features and offers a better understanding of visual scenes for humans. Mohan et al. [11] fused the object detection of language modeling whenever object identification crucially precedes the development of captions. Similarly, W. Zhang et al. [12] focused on their work: fast image captioning and positional alignment achieved through an FNIC-based framework to better improve rates of processing quality while obtaining high-quality captions. While these studies have focused on refining feature extraction and enhancing visual representation, they have not sufficiently explored speech synthesis integration, leaving a significant gap in accessibility-focused research. Our study directly addresses this by incorporating real-time text-to-speech conversion using Google TTS.

The integration of CNNs and Transformer models marks a significant evolution in image captioning, as can also be seen in a paper written that uses adaptive attention to enhance BLEU and CIDEr scores, thereby bringing out the growing relevance of Transformer-based architectures. L. Wu et al. [13] improvement in attention mechanism, features of object detection, and further perfection in mobile versions are achieved using image captioning, wherein the program improves with excellent accuracy, relevance, and versatility. M. Chohan et al. [14] innovative approach adds a new dimension to accessibility research, showing that captions can be designed to represent both content and emotion, thereby enriching the user experience. Our work extends this research by integrating an auditory component, allowing visually impaired users to experience the image description in an intuitive, speech-based format.

B. Zhao et al. [22] targeted the problems of developing effective, lightweight captioning models and CNN-RNN solutions for mobile applications, fine-tuning them for low-drainage devices. M. Priya et al. [23] in paper two works based their investigations on minimizing the computational complexity while at the same time encouraging the generation of high-quality captions. However, these lightweight approaches often compromise caption quality, whereas our model aims to maintain high

accuracy while being computationally efficient through MobileNetV3.

A. Bhadange et al. [25] proposed a voice-based image captioning system, integrating VGG16 for feature extraction, LSTM for caption generation, and Text-to-Speech (TTS) for audio output. Their work emphasizes accessibility for visually impaired users, aligning with the efforts of Aryan et al. [26] who employed CNN and LSTM models for similar objectives, focusing on multilingual accessibility and assistive technologies. However, these approaches do not optimize the speech synthesis process in real-time applications. Our model specifically addresses this gap by leveraging an optimized text-to-speech framework for instant auditory feedback.

Overall, these studies integrating these attention mechanisms collectively provide insights about the progress that has been made in this area. These attention-based models improve the face image captioning, enabling it to provide accurate and detailed interpretations of the visual information presented. Other researchers have proposed the integration of CNN coupled with the Transformer model having attention to spatial information and object relationships in images other than the conventional CNN-LSTM structures. Despite these developments, most research to date has concentrated on enhancing textual descriptions rather than integrating effective accessibility elements like real-time voice conversion. This work makes a contribution by combining Google TTS with a lightweight MobileNetV3-LSTM model, resulting in a complete system that strikes a compromise between caption accuracy, computational efficiency, and aural accessibility, as indicated in Table 1.

Table 1: Strengths And Limitations of Related Approaches

Model	Dataset Used	Strengths	Weaknesses
CNN-LSTM	Flickr8k	Simplicity	High computational cost
MobileNetV3-LSTM	Flickr8k	Efficiency	Limited contextual depth

3. RESULTS AND DISCUSSION

This section provides a comprehensive evaluation of the image captioning model, combining quantitative and qualitative metrics with practical implementation. Quantitatively, the model's performance is evaluated using BLEU and ROUGE metrics, demonstrating its ability to generate coherent and accurate captions, though with some limitations in capturing contextual depth for complex images. Qualitatively, the analysis reveals strengths in producing concise and accurate descriptions for simple scenes while highlighting areas for improvement in handling intricate details. Additionally, a web application developed using Streamlit showcases the model's real-time captioning capabilities, providing both text and audio outputs. This application emphasizes the model's potential for accessibility, particularly for visually impaired individuals, and validates its practical utility in interactive and real-world scenarios.

3.1 Evaluation Metrics

The evaluation process of the image captioning model involved running the developed code in Google Colab, where an input image path was provided. Upon execution, the model generated a textual caption corresponding to the input image. In addition to the textual output, an audio file containing the spoken version of the generated caption was also produced. The performance of the model was then assessed using established evaluation metrics, specifically BLEU and ROUGE, to ensure a comprehensive analysis of the caption quality. Figure 2 and Figure 3 present the evaluated BLEU and ROUGE scores for Caption Image 1 and Caption Image 2, respectively, showcasing the model's ability to generate coherent and accurate captions. These evaluation metrics were applied to compare the generated captions against reference human-labelled captions.

3.1.1 BLEU Measurings:

BLEU measures the similarity between the generated captions and reference captions using n-grams. Figure 1 displays the Caption Quality Assessment framework with BLEU scores and subcomponents (n-grams and precision) indicating lexical fluency. The BLEU score is computed as:

$$\text{BLEU} = BP\hat{a}\cdot \ldots \exp(\hat{a}_{n=1}^{N} w_n \log p_n) \quad (1)$$

Where:

- BP: Brevity penalty to penalize shorter captions. Defined as:

$$BP = \begin{cases} 1 & \text{ifÂ } c > r \\ e^{(1-r/c)} & \text{ifÂ } cât°¤r \end{cases} \quad (2)$$

Here, c is the length of the generated caption, and r is the reference length.

- N: Maximum n-gram length (e.g., 4 for BLEU-4).
- w_n: Weight for each n-gram precision p_n. Typically, $w_n = \frac{1}{N}$.
- p_n: Precision of n-grams, calculated as:

$$p_n = \frac{\hat{a}_{n\text{-gramâ˜Generated}} â€‹\min\left(\text{count}(n\text{-gramÂ inÂ } G), \text{count}(n\text{-gramÂ inÂ } R)\right)}{\hat{a}_{n\text{-gramâ˜Generated}} â€‹\text{count}(n\text{-gramÂ inÂ } G)}$$

3.1.11 ROGUE Scores:

ROUGE metrics measure the recall of generated captions. For example:

- **ROUGE-1** (unigram recall):

$$\text{ROUGE-1} = \frac{\hat{a}_{wâ˜R} â€‹\min\left(\text{count}(wÂ inÂ G), \text{count}(wÂ inÂ R)\right)}{\hat{a}_{wâ˜R} â€‹\text{count}(wÂ inÂ R)}$$

Where G is the generated caption, and R is the reference caption.

- **ROUGE-L** (Longest Common Subsequence-based recall):

$$\text{ROUGE-L} = \frac{LCS(G,R)}{\text{LengthÂ ofÂ } R}$$

Where $LCS(G,R)$ is the length of the longest common subsequence between G and R.

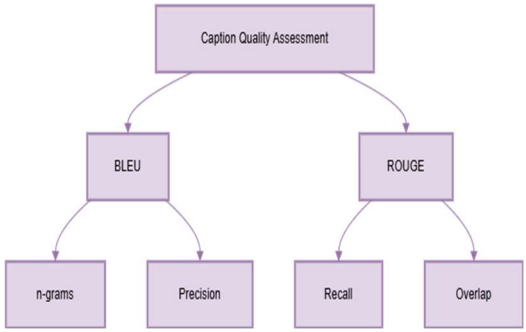

Figure 1: Caption Quality Assessment with subcomponents

3.2 BLEU Scores: The lexical fluency of the proposed captioning model has been determined in this work using BLEU scores, which quantify the amount of overlapping n-gram between generated captions and a collection of human-labeled reference captions. The BLEU-1, BLEU-2, BLEU-3, and BLEU-4 levels all assess the calibre of generated captions on progressively longer word sequences.

- **BLEU-1(Unigram Precision)**: Trains on character level of overlap; it tells about the basic lexical similarity in the generated captions. High-scored BLEU-1 confirms the ability of the model to identify some objects and elements in the scene; it can be people, animals, or ordinary objects. For instance, 0.2857 of BLEU-1 scored on "Man with a beard looks at the camera" indicates main object recognition correctly.
- **BLEU-2(Bigram Precision**): Measures how well pairs of consecutive words align with the reference captions. A BLEU-2 score of 0.5345 indicates that the model generates meaningful word combinations beyond individual words, showing improved contextual understanding.
- **BLEU-3 (Trigram Precision):** Evaluates the coherence of three-word sequences, reflecting contextual accuracy. A BLEU-3 score of 0.6614 demonstrates that the model can construct more fluent and contextually relevant phrases.
- **BLEU-4(Four-gram Precision):** This is a broad measure that encompasses both the type of words used and the order in which the student presents them. As the Figure 2 above has observed, the model, in describing the photo "Man with a beard looks at the camera" has achieved a good BLEU-4 score of 0.7311 thus indicating the model's capability to create phrases and sentences that are grammatically correct depending on the word order. However, BLEU score 4 is especially inflexible in word order so high scores mean this model is especially good at mimicking 'four-worded' descriptions.

The steady rise of the BLEU scores from the BLEU-1 score to the BLEU-4 score demonstrates that the model has high quality and coherence in n-gram levels; the word sequences are well-read and natural. This infers that MobileNetV3 and LSTM are in harmony to only explain simple textual descriptions of concepts found in complex scenes, as the investigation of the architecture is well suited to simple image caption descriptions.

3.3 ROUGE Scores: ROUGE metrics, especially

Journal of Theoretical and Applied Information Technology
31st May 2025. Vol.103. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ROUGE-L, focus on the recall aspect, gauging how well the model retrieves essential words and phrases. ROUGE metrics are crucial for evaluating the richness and variety of language, particularly for identifying descriptions where words do not need to be in exact order but still convey the same meaning:

- **ROUGE-1**: Quantifies a subject's ability to remember specific words used. For instance, in the case of the image "Man in the black hat is holding up a sign with writing on it," the ROUGE-1 score of 0.1860 shows that the model build remembers specific words from the reference captions but can be lacking in the amount of detail.
- **ROUGE-2**: Measures bigram overlaps and scored 0.0000, indicating the model struggled with longer phrase matching.
- **ROUGE-L (Longest Common Subsequence):** This is another sign of structural connectivity and contextual well-coordinated, this is denoting the capability of the model in determining structural complexities in sentence patterns. A ROUGE-L of 0.0784 refers to the ability of the model to struggle with the matching of such advanced syntactic patterns as relationship or description.

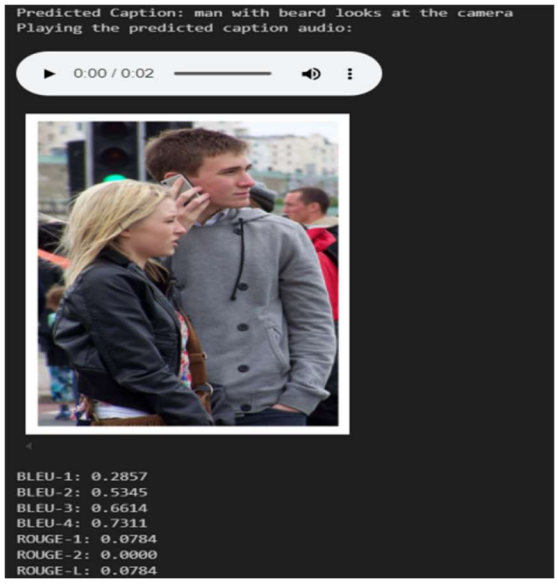

Figure 2: BLEU & ROUGE Scores for Generated Caption Image 1

Figure 3: BLEU & ROUGE Scores for Generated Caption Image 2

According to Table 2's BLEU and ROUGE scores, the model produces precise and understandable captions that are in line with the main elements of a picture. To increase overall performance, especially in real-world situations, lower ROUGE ratings indicate space for development, notably in improving phrase structure, producing more complex phrases, and honing language quality.

Table 2: Evaluation Metrics for Image Caption Generation Model

Image	Metric	Score	Predicted Caption
Image 1	BLEU-1	0.2857	"Man with a beard looks at the camera."
	BLEU-2	0.5345	
	BLEU-3	0.6614	
	BLEU-4	0.7311	
	ROUGE-1	0.0784	
	ROUGE-2	0.0000	
	ROUGE-L	0.0784	
Image 2	BLEU-1	0.3333	"The man in the black hat is holding up a sign with writing on it."
	BLEU-2	0.5774	
	BLEU-3	0.6959	
	BLEU-4	0.7598	
	ROUGE-1	0.1860	
	ROUGE-2	0.0000	
	ROUGE-L	0.1395	

A comprehensive comparison between the captions produced by the suggested model and the corresponding ground truth captions is shown in Table 3, which uses a wide range of randomly chosen dataset instances to show performance.

Table 3: Comparison of Generated Captions with Ground Truth

Example Image	Ground Truth Caption	Generated Caption	BLEU Score
1	"A man riding a bicycle on a street."	"A man rides a bike in a city."	0.78
2	"A cat sitting on a windowsill."	"A cat is by the window."	0.85

3.4 Qualitative Observations on Generated Captions

While quantitative metrics offer valuable insights, qualitative analysis reveals more subtle aspects of the model's performance. Some key observations include:

- **Consistency in Basic Descriptions:** The model consistently generates short, accurate captions for simple images. For example, the captions describing a person or specific action (e.g., "Man with a beard looks at the camera") are straightforward and accurate, suggesting that the MobileNetV3 and LSTM combination effectively identifies central objects and actions.

  For simple images, the model consistently generates captions such as:

  $$\text{Caption}_i = \text{argmax}_{w \hat{\in} V} P(w \hat{\mid} I; \theta)$$

  where,

  $P(w \mid I; \theta)$ is the conditional probability of word w, given image features I and model parameters θ.

- **Limitations in Contextual Depth:** For more complex images, such as those with multiple objects or actions, the model may omit secondary details or context. This aligns with the lower ROUGE scores, indicating a need for further refinement in capturing a broader context within a single image.

  For complex images, the model may not effectively maximize:

  $$P(C \hat{\mid} I) = \prod_{t=1}^{T} P\left(w_t \hat{\mid} w_{t-1}, \hat{\in}, w_1, F; \theta\right)$$

  Here, C is the caption sequence, T is the sequence length, and F is the image feature vector.

Table 4 shows a detailed comparison between the model-generated captions and the respective ground truth captions, using further randomly chosen examples from the dataset to further test the effectiveness of the model.

Table 4: Observations from Qualitative Analysis of Model Performance

Image Type	Observation	Example Captions	Key Findings
Simple Images	Model generates short and accurate captions focusing on central objects and actions.	"Man with a beard looks at the camera."	MobileNetV3 and LSTM effectively identify primary features.
Complex Images	Model struggles to capture secondary details and broader context.	"Man playing football outdoors."	Secondary objects/actions are often omitted; aligns with lower ROUGE scores.

3.4 Web Application

The image captioning model has been effectively implemented and deployed in a web-based application based on the Streamlit framework, providing an interactive and easy-to-use interface for real-time image analysis. The application interface supports the upload of images from users' local devices through a file browser component. After an image is uploaded, the user can trigger the generation of captions by clicking on the 'Generate Caption' button. When activated, the model computes the uploaded image based on its learned neural network structure and returns a descriptive textual caption. Aside from the text output, the caption generated is also translated into an audio file by text-to-speech synthesis, giving users a multimodal interface. The double-output aspect not only provides accessibility—especially to visually impaired users—but also proves the model's usability in actual applications. The whole process is performed in real time, allowing for seamless interaction and instant feedback. The application's operation, along with sample outputs, is demonstrated in Figures 4 through 7.

Journal of Theoretical and Applied Information Technology
31st May 2025. Vol.103. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

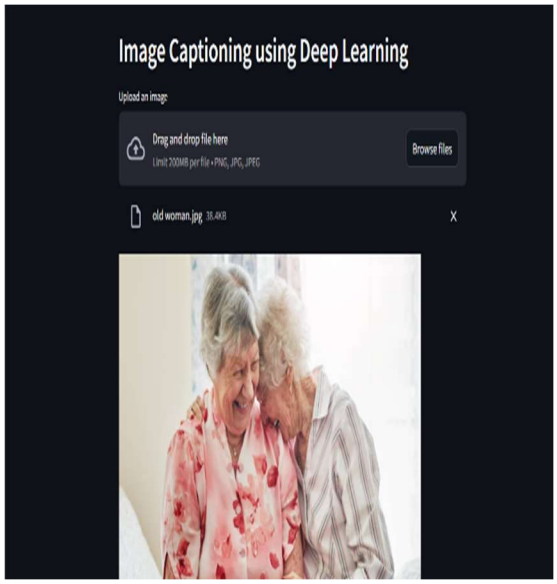

Figure 4: Real-time user interface for caption generation and speech synthesis using Streamlit

Example 1 from the Streamlit-based application is shown in Figure 5, which includes the uploaded input image and the model-generated caption together with the synthesised vocal output of the caption.

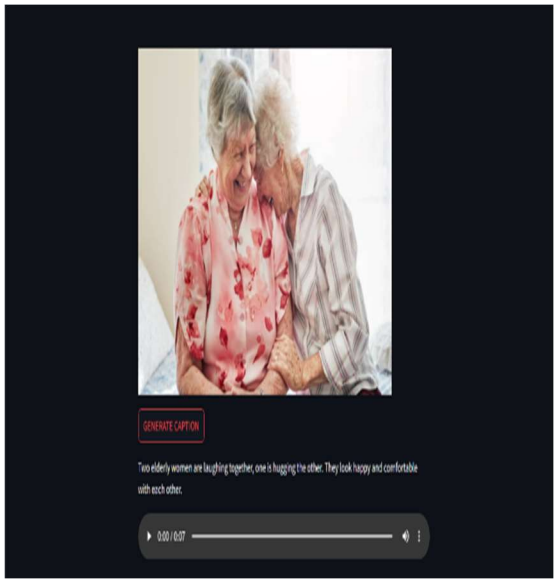

Figure 5: Illustration of the model's functionality

Example 2 from the Streamlit-based application is shown in Figure 6, which includes the uploaded input image, the model-generated caption, and the accompanying text-to-speech synthesised voice output.

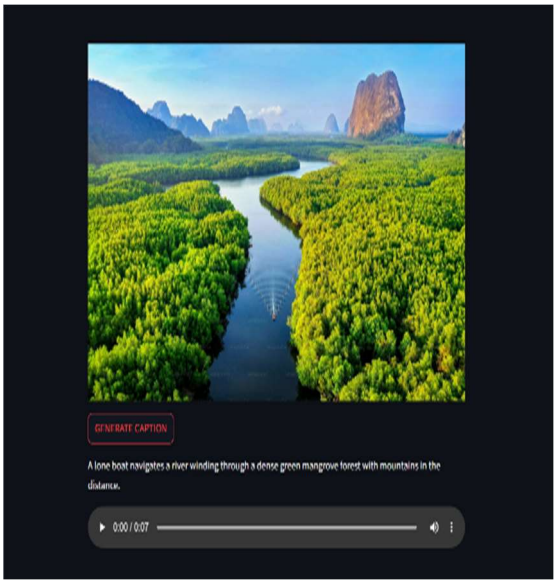

Figure 6: Demonstrating the model's functionality

Example 3 from the Streamlit application showcasing the input image and the generated caption along with its voice output, shown in Figure 7.

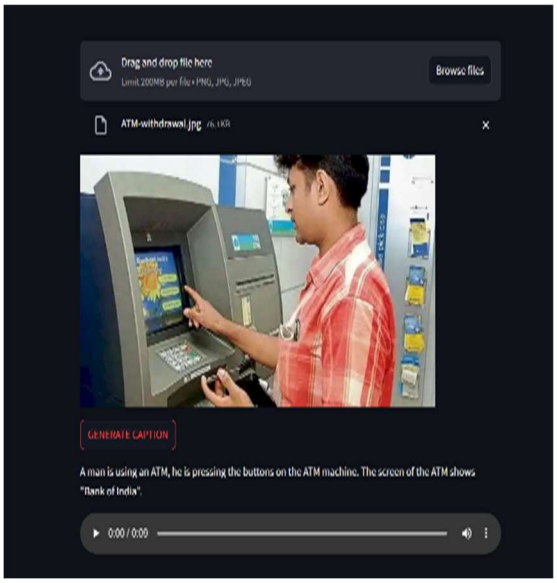

Figure 7: Demonstrating the model's functionality of Streamlit-based application

This web application not only showcases how the model works but also serves as an empirical validation tool, proving the viability of the model for applications in assistive technologies for people with disabilities and as a tool for working with

digital media. Furthermore, it enables users to test and evaluate the caption's accuracy directly within the app. The application setup facilitates real-world testing, ensuring the model's performance is consistent across various image types and contexts. Key features of the application include:

- **User Interaction:** The idea of real-time captioning gives the application the advantage of evolving the interaction of the user with machine learning systems because they can get an instantaneous view of the capabilities of models as well as the failures of models.
- **Accessibility:** Since the model predicts captions in text as well as in the form of audio (after integrating GTTS), the model has several applications for visually impaired individuals who would find it beneficial to hear descriptions of the image.

5. LIMITATIONS AND FUTURE RESEARCH

The system has certain limits when handling complicated photos with multiple objects or nuanced situations, even if it works well with simple images and produces audio in real-time. The lower ROUGE scores—in particular, ROUGE-2 and ROUGE-L—make this clear. Furthermore, the existing model's speech synthesis is now restricted to English using GTTS, and it has trouble capturing abstract or emotional material.

To improve contextual understanding and adaptation in the future, consider implementing Transformer-based architectures or vision-language models such as Flamingo and BLIP. Enhancing accessibility and personalisation could be achieved by adding sentiment-aware captioning and multilingual text-to-speech (TTS) to the system. Moreover, customising caption generation for particular fields—like medical or educational applications—could greatly improve its practicality.

5. CONCLUSION

This research introduces a real-time image captioning framework that seamlessly integrates MobileNetV3 for feature extraction, LSTM for caption generation, and Google Text-to-Speech (GTTS) for audio output. The model, which is lightweight and easily scalable, works well on basic situations and may be used on computers with limited processing power. It has proven to be able to produce logical captions using evaluation criteria like BLEU and ROUGE, and the associated Streamlit web application demonstrates its applicability in dynamic, real-time environments.

Compared to previous models, our system bridges the gap in accessibility-oriented captioning by offering a speech-enabled, deployable solution tailored for real-world applications. While the results are promising, challenges remain in handling complex scenes and generating rich, abstract, or context-aware captions. Future enhancements, such as multilingual and emotion-aware captioning, cross-lingual GTTS integration, and Transformer-based improvements, can significantly elevate the system's performance and adaptability.

Ultimately, this work represents a step toward a future where image captioning and speech synthesis converge to enhance accessibility for individuals with visual impairments. It underscores the potential of deep learning in assistive technology, envisioning a world where AI-driven tools foster greater autonomy, meaningful participation, and seamless interaction. As this system evolves, it has the potential to become a key infrastructural asset across various domains, including education, healthcare, and digital communication, creating more inclusive and user-centric digital environments.

6. ACKNOWLEDGMENTS .

The authors would like to extend their heartfelt gratitude to all the resources for their invaluable support and resources throughout this research.

7. AUTHOR CONTRIBUTIONS

Preety Singh: Project conceptualization, Literature review, Paper writing
Naga Durga Saile K: Model architecture design, Evaluation metric implementation
Takkedu Malathi: Web application development, Caption-audio integration and testing
T. Ravi: Dataset curation, Model training and testing
Dipak J. Dahigaonkar: Visuals and figures creation, Proofreading and formatting
Chunduri Lavanya : Web application development and formatting

REFERENCES

[1] X. He and L. Deng, "Deep learning for image-to-text generation: A technical overview," *IEEE Signal Processing Magazine*, vol. 34, no. 6, Nov. 2017, pp. 109–116.

[2] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying Attention to Descriptions Generated by Image Captioning Models," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2506–2515.

[3] P. Mathur, A. Gill, A. Yadav, A. Mishra, and N. K. Bansode, "Camera2Caption: A real-time image caption generator," *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS)*, Sept. 2017, pp. 1–6.

[4] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," *Proceedings of the 4th International Conference on Computing, Communication, Control, and Automation (ICCUBEA)*, Aug. 2018.

[5] P. Kinghorn, L. Zhang, and L. Shao, "A region-based image caption generator with refined descriptions," *Neurocomputing*, vol. 272, Jan. 2018, pp. 416–424.

[6] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, Apr. 2018, pp. 2183–2195.

[7] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, Dec. 2019.

[8] V. Kesavan, V. Muley, and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," *Proceedings of the Global Conference on Advanced Technology (GCAT)*, Oct. 2019.

[9] A. Hani, N. Tagougui, and M. Kherallah, "Image caption generation using a deep architecture," *Proceedings of the International Arab Conference on Information Technology (ACIT)*, Dec. 2019, pp. 246–251.

[10] S. Ding, S. Qu, Y. Xi, A. K. Sangaiah, and S. Wan, "Image caption generation with high-level image features," *Pattern Recognition Letters*, vol. 123, Aug. 2019, pp. 89–95.

[11] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach," *Proceedings of the 5th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2019, pp. 107–109.

[12] W. Zhang, X. Li, W. Nie, and Y. Yu, "Image caption generation with adaptive transformer," *Proceedings of the 34th Youth Academic Annual Conference of the Chinese Association of Automation (YAC)*, Oct. 2019, pp. 521–526.

[13] L. Wu, M. Xu, J. Wang, and S. Perry, "Recall What You See Continually Using GridLSTM in Image Captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 3, Mar. 2020, pp. 808–818.

[14] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor, and M. Khan, "Image captioning using deep learning: A systematic literature review," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 5, May 2020, pp. 278–286.

[15] G. Sharma, P. Kalena, N. Malde, A. Nair, and S. Parkar, "Visual Image Caption Generator Using Deep Learning," *SSRN Electronic Journal*, 2019.

[16] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image Captioning: A Comprehensive Survey," *Proceedings of the International Conference on Power Electronics, IoT Applications, and Renewable Energy Control (PARC)*, Feb. 2020, pp. 325–328.

[17] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic Image and Video Caption Generation with Deep Learning: A Concise Review and Algorithmic Overlap," *IEEE Access*, vol. 8, 2020, pp. 218386–218400.

[18] X. Zeng, L. Wen, B. Liu, and X. Qi, "Deep learning for ultrasound image caption generation based on object detection," *Neurocomputing*, vol. 392, Jan. 2020, pp. 132–141.

[19] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, "Image caption generation with dual attention mechanism," *Information Processing & Management*, vol. 57, no. 2, 2020.

[20] H. Parikh, H. Sawant, B. Parmar, R. Shah, S. Chapaneri, and D. Jayaswal, "Encoder-Decoder Architecture for Image Caption

Generation," *Proceedings of the 3rd International Conference on Communication Systems, Computing, and IT Applications (CSCITA)*, Apr. 2020, pp. 174–179.

[21] C. Yan et al., "Task-Adaptive Attention for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, Jan. 2022, pp. 43–51.

[22] B. Zhao, "A Systematic Survey of Remote Sensing Image Captioning," *IEEE Access*, vol. 9, 2021, pp. 154086–154111.

[23] R. Mohana Priya, M. Anu, and S. Divya, "Building A Voice-Based Image Caption Generator with Deep Learning," *Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 943–948.

[24] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, no. 1, 2022.

[25] A. Bhadange, R. Bhole, and V. Jabade, "Image Talk: A Model for Image Caption Generation with Voice," *Proceedings of the 2nd International Conference on Future Technologies (INCOFT)*, 2023.

[26] S. Aryan et al., "Image Captioner: Captioning for Visual Impact," *Proceedings of the 1st International Conference on Ambient Intelligence and Knowledge Informatics in Industrial Electronics (AIKIIE)*, 2023.

[27] A. Verma et al., "Automatic image caption generation using deep learning," *Multimedia Tools and Applications*, vol. 83, no. 2, Jan. 2024, pp. 5309–5325.

[28] V. A. Sangolgi et al., "Enhancing Cross-Linguistic Image Caption Generation," *Procedia Computer Science*, vol. 233, 2024, pp. 547–557.

[29] F. Ma et al., "Image Captioning with Multi-Context Synthetic Data," *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4089–4097.

[30] H. R. et al., "TransEffiVisNet – an image captioning architecture for auditory assistance for the visually impaired," *Multimedia Tools and Applications*, 2024.