Beyond Sight: Generating Spoken Descriptions

Renuka Bhandari[1], Nidhi Yadav[2], Preeti Warrier[3]
[1,2,3]*Department of ENTC Engineering, Army Institute of Technology (AIT), Pune, Maharashtra, India, 411047*

***Abstract*—Image captioning is a task in computer vision that requires creating detailed descriptions for images. This process links visual content with natural language, allowing machines to interpret and explain visual scenes. This study presents a sophisticated system that employs a pre-trained convolutional neural network (CNN) to obtain comprehensive features from images. These features are integrated with an attention mechanism and are utilized to generate captions using a recurrent neural network (RNN).**

To develop thorough feature vectors for images, several pre-trained convolutional neural networks, such as Inception V3, were used in a planned and effective manner. The process of decoding makes use of the Long Short-Term Memory (LSTM) model, which was chosen because it is effective at creating clear and descriptive sentences. To further enhance performance, an innovative integration of the Inception V3 attention model was implemented, allowing the system to concentrate on specific areas of the image during the learning process. Experimental results from tests on the flickr8k dataset show very good performance, similar to the best current methods.

The main goal of the study is to help people with visual impairments by giving them a way to learn about visual information through sound. Traditional approaches, like image captions, are not sufficient to address the specific requirements of the visually impaired community. Therefore, this study tackles the pressing need for an advanced solution—a model that is not only capable of analyzing images but also able to convert them into spoken descriptions. The description is created by using the Google Text-To-Speech (gTTS) API. This innovative method seeks to connect visual content with auditory comprehension, paving the way for a more inclusive and accessible future.

***Index Terms*—Machine Learning, Computer Vision, Image Captioning, Deep Learning, Feature Extraction, Long Short-Term Memory (LSTM), Google Text-To-Speech (GTTS)**

I. INTRODUCTION

In today's world, where visual content is prevalent, people with vision impairments often find it very challenging to access and understand images. This study aims to provide a thorough solution by introducing a deep learning model that generates spoken descriptions of images. The primary issue that this study aims to address is how to allow people who are visually impaired to perceive and understand visual information through the sense of hearing. For the blind, standard techniques like picture captions are inadequate. Therefore, a solution is needed that connects visual information with auditory understanding—one that goes beyond analyzing images by also transforming them into spoken explanations.

This work is centered on utilizing a deep learning model, specifically an LSTM with an attention block, to examine and analyze photographs. The model is trained using a broad range of photographs that come with corresponding descriptions from the chosen dataset, Flickr8K.

Metrics such as the METEOR score, the BLEU score, and evaluations by human judges will be used to measure how well the model performs. These metrics will evaluate how well the generated captions match human-like descriptions in terms of both clarity and accuracy.

II. LITERATURE SURVEY

The environment for technological accessibility has been continuously evolving, showing steady progress in supporting diversity. Braille and screen readers are examples of earlier inventions that laid the foundation for today's efforts to make information available to people who are blind. However, the issue continued to exist within the area of visual information. Textual

explanations were included with photographs as a way to address the issue, showing an effort to help blind individuals understand the visual elements presented [1]. Despite this, the method turned out to be very time-consuming and difficult to scale up [2][3]. Therefore, additional efforts were made to enhance the clarity of the visual elements for individuals with visual impairments. The purpose of this research is to incorporate artificial intelligence into the creation of visual content that can be accessed by people with visual impairments, by examining existing research and understanding the progress and challenges in this area.

A. Image Captioning Techniques

Numerous research studies have looked into various approaches for generating descriptions of images. Several studies have focused on the application of RNNs for generating sequential data and CNNs for feature extraction from images [4]. The enhancement of the coherence and relevance of produced captions has highlighted the importance of attention processes.

B. Attention Mechanisms in Deep Learning

Numerous research efforts have focused on how attention mechanisms are incorporated into deep learning models[5]. When making captions, attention mechanisms enable the model to focus on specific parts of an image, leading to more detailed and thorough descriptions overall. The popularity of the attention model is due to its capability to focus on specific visual areas dynamically and enhance the accuracy of descriptions. The attention mechanism, which works like the human visual system [6], enables focusing on important parts selectively, unlike uniform methods such as the graph-based approach. More complex scenes benefit from this interpretability, which also results in more detailed and contextually appropriate captions[7][8]. The attention model offers an advanced and context-sensitive approach to generating image captions by assigning varying levels of importance to different parts of an image, thereby enhancing the system's ability to understand and represent complex relationships within the scene.

C. Datasets for Image Captioning

Besides the Flickr8K dataset mentioned in the problem statement [9], image captioning studies have used many other datasets like Flickr30k, VIST (Visual Storytelling), SBU (Socially-embedded Image Dataset), ADE20K (ADE20K Scene Parsing), and MS COCO (Common Objects in Context). It's important to understand the pros and cons of different datasets so you can build a trustworthy model.

D. Evaluation Metrics

The effectiveness of picture captioning models has been measured using BLEU (Bilingual Evaluation Understudy) score[10], METEOR score[11], and other similar metrics. It's important to look at how these metrics check the quality of the captions that are made in order to assess the research outcomes.

E. Text-to-Speech Technologies

For the final part of the research, studying text-to-speech technology is very important. The research's success will rely on how effectively different TTS libraries, like Nancy, TWEB, M-AI-Labs, LibriTTS, and LJ Speech[13], convert the generated captions into speech that sounds natural.
After looking closely at the different TTS tools available[14] [16], we decided that using Google Text-to-Speech along with LSTM networks works best for our research. The choice was made because LSTM showed very good results within Recurrent Neural Networks, which fit exactly with what our research needed and aimed to achieve. Using GTTS along with LSTM helps produce high-quality voice synthesis and makes the caption creation process more efficient overall.

III. METHODOLOGY

Creating detailed descriptions of what is seen in images is called image captioning, and it is a difficult task that combines computer vision with natural language processing. Convolutional Neural Networks (CNN) are important for image recognition because they help find detailed information in images and understand what they look like. These features are then given to Recurrent Neural Networks (RNN), which are really good at handling data that comes in a sequence and can keep track of a lot of specific information from the context. RNNs create captions that make sense and fit well with the situation by using the information from the words they have already generated.

We used a powerful set of tools to help our picture captioning research work as well as it could. CNNs, RNNs, and the attention mechanism were all easily included in the flexible deep learning system provided by TensorFlow and Keras[12]. Pandas and NumPy helped handle data more efficiently, which made it easier to integrate the Flickr8k dataset. By improving how it analyzes language, NLTK (Natural Language Toolkit) made the quality of output captions better. Our research managed to combine advanced deep learning techniques with dependable data processing, which was made possible by using tools like Tensor Flow [12], Keras, NumPy, Pandas, NLTK, and Flickr8k. This helped our model that focuses on attention to create good captions for images.

Adding attention mechanisms helps make the image captioning process better. The attention methods, which became popular in machine translation tasks first, let the model create each word of the caption in Fig.1 by focusing on specific parts of the image. This is like how the human brain works, where important parts of a picture are focused on when someone describes it aloud. When people look at photos, they usually pay attention to the things that stand out the most. Attention-based processes function in a similar way, helping the model to "focus on" specific areas during training.

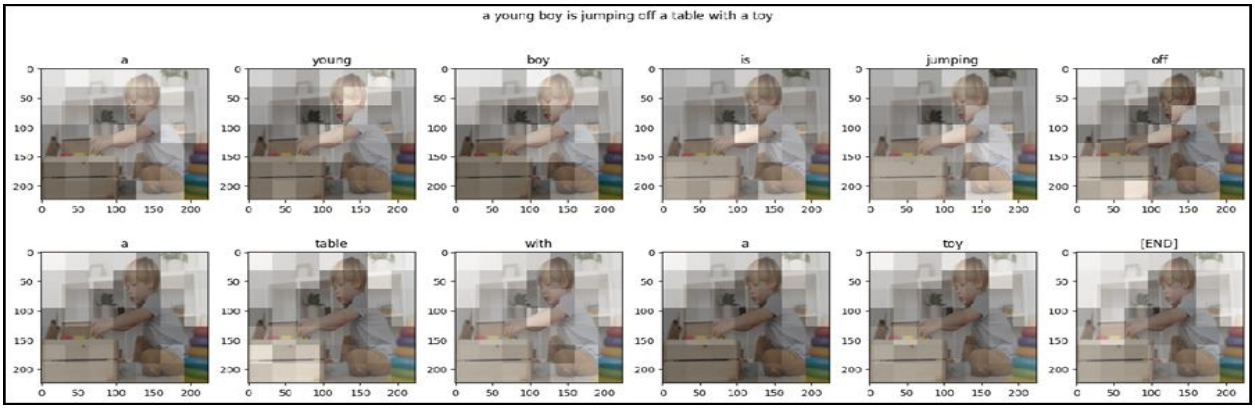

Fig. 1. Attention mechanisms, selectively focus on specific regions of the image while generating each word of the caption, Example of caption generation in between epoch completion giving almost accurate but not correct caption

The model's ability to create detailed and context-sensitive descriptions for different images is enhanced by combining CNNs, RNNs, and attention mechanisms.

The attention mechanism can be implemented in two different methods. In the case of soft attention, as shown in Fig.2 A, the model focuses on weighted image features rather than using the entire image as input to the long short-term memory (LSTM). Soft attention works by reducing the importance of irrelevant regions through the use of a low weight, which is applied by multiplying it with the relevant feature map. It is simple and easy to calculate. Whereas hard attention mechanisms, as shown in Fig.2 B, use a stochastic sampling model for their functioning. Sampling is used to guarantee precise gradient descent during the back propagation process, and the results are then combined using the Monte Carlo method. Monte Carlo methods run complete episodes from start to finish to calculate an average based on all the results gathered through sampling. The accuracy of hard attention depends on both the amount of samples used and how well the sampling is done. However, some important areas are often missed or forgotten because the attention model processes a long sequence of data step by step. We are using the Long Short-Term Memory (LSTM) - RNN architecture to solve this problem. Long Short-Term Memory, or LSTM, is a specialized type of Recurrent Neural Network designed to address the limitations of traditional RNNs in handling long-term dependencies in sequential data. Long Short-Term Memory networks differ from standard Recurrent Neural Networks because they include special mechanisms called gates and memory cells. These components allow the network to selectively remember or forget

information over long sequences, making it more effective at handling complex patterns in data.

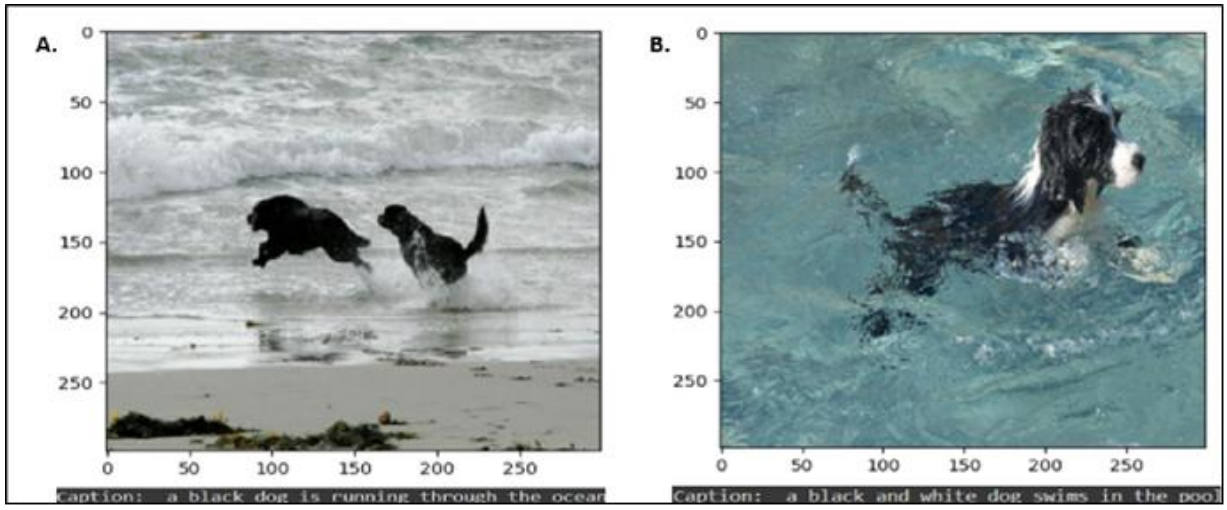

Fig. 2 . Precise and correct caption generated after model training completion of a dog playing in the water. Caption for image A is generated through soft attention model while image B is generated through hard attention model

For tasks that involve data arranged in a sequence, such as language modelling and predicting time series, LSTMs are particularly good at capturing and retaining important contextual information across different time steps. This is achieved through the use of input, forget, and output gates, which enable the regulated movement of information into and out of the memory cell.

A good method for describing images involves using Convolutional Neural Networks (CNNs) along with attention mechanisms in combination with LSTM-RNN structures. LSTM enhances the model's ability to understand the sequence of connections in the output descriptions.

While CNN is effective at capturing spatial features from images, the attention mechanism focuses on specific areas of interest. This combined approach uses the unique strengths of each part to create a full model that can handle tasks needing both order and spatial awareness. This collaborative method significantly enhances overall effectiveness, producing outcomes that are more detailed in context and more precise.

The development of an image captioning model involves thoughtful attention to the steps of data extraction and processing. Datasets like Flickr8k are great for evaluation and training because they offer a wide range of images along with detailed explanations for each. To ensure seamless integration of the data into the training process, the data preparation involves cleaning and organizing the information [4][5]. To enhance its ability to generate accurate and meaningful captions as shown in Fig.[2], the algorithm is trained to recognize patterns and connections between visual features and corresponding textual information.

This approach can be applied to various other areas and is not limited to generating descriptions for images. The attention mechanism was first created for machine translation but has since been applied to many different tasks in natural language processing. Tasks like text summarizing show how versatile the model is, as it focuses on key information and creates clear, concise summaries. In addition, attention mechanisms have been demonstrated to enhance the clarity and effectiveness of models when used in tasks such as answering questions and analyzing sentiment.

In summary, the processes of natural language processing, attention mechanisms, and image recognition work together in a delicate way during the captioning process. By combining attention mechanisms with CNNs and RNNs, models can effectively transform visual content into clear and contextually relevant text descriptions. The ability of these systems is further improved through the use of datasets, precise data extraction, and effective model

training, which opens up more possibilities for their application across different areas.

IV. RESULTS AND DISCUSSION

The image captioning system, developed by integrating advanced Recurrent and Convolutional neural networks with a detailed attention mechanism, achieved very promising results. The model successfully generated clear and logical text descriptions by paying attention to the main features and detailed elements within the images. Thanks to the important and detailed role that the attention mechanism played. Over the long training phase, the system consistently demonstrated a clear pattern of improvement, with significant increases observed in both accuracy and loss metrics.

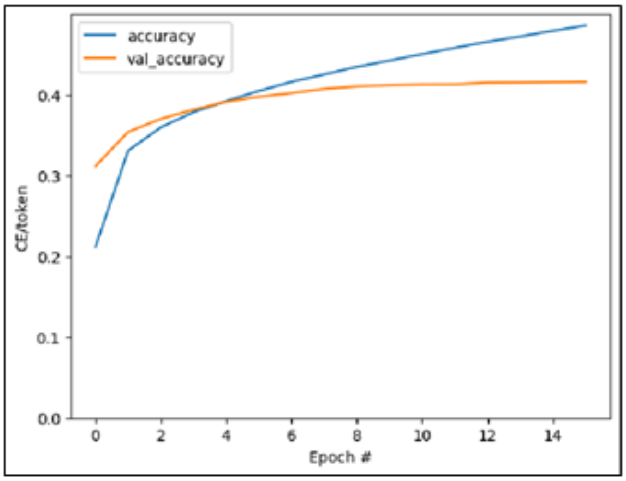

Fig. 3. The above graph compares the expected result and the obtained accuracy over time as epoch overlaps for caption generation.

The accuracy trend shown in Fig. 3 clearly shows that the model is becoming better at generating captions that are highly responsive to small details in the visual content and closely match the annotations provided as the correct answers. At the same time, the significant reduction in loss, as shown in Fig.4,

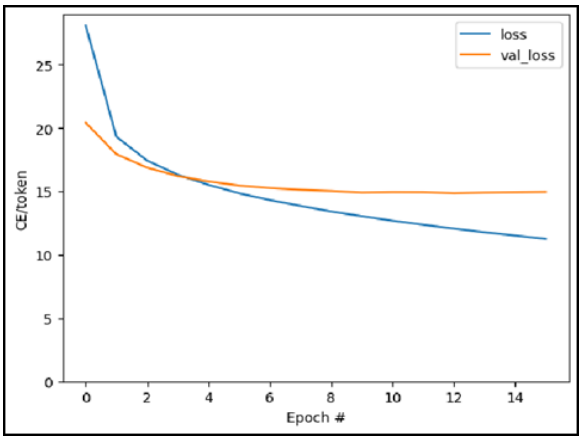

Fig. 4. This graph shows the loss of error between the captions generated and calculated captions over time as the epoch overlaps clearly shows the model's ability to minimize the differences between expected and actual captions, which indicates its strong learning abilities. The combination of these attention processes along with thorough model training led to an image captioning system that is highly reliable, capable of providing meaningful text descriptions for a wide range of visual content with greater accuracy. As shown in Fig.2, the model's captions were not only highly accurate but also effectively conveyed the complex visual characteristics of the images, highlighting the system's capability to understand and represent the detailed aspects of the visual content.

V. CONCLUSION

Our initiative in generating image captions marks a major progress in the areas of computer vision and natural language processing. By combining convolutional neural networks (CNNs), long short-term memory recurrent neural networks (LSTM-RNNs), and attention mechanisms, we have introduced a new approach and developed a robust framework. This process is very effective at creating written descriptions that are. Both naturally cohesive and richly contextualized for a wide range of varied pictures. Choosing to incorporate the Flickr8k dataset deliberately significantly broadened our model's experience with diverse linguistic and visual contexts. Fig. 2 demonstrates this strategic enhancement, significantly improving the overall performance and adaptability of our picture captioning system.

A notable enhancement to our work is the inclusion of a text-to-speech feature, which represents an important advancement that enhances the system's usability and makes it more accessible to users. The impact of our work is enhanced through the creative integration of GTTS, which also increases its inclusivity and adaptability to the diverse needs and preferences of users.

This study emphasizes the broader effects of attention mechanisms and also contributes to advancing the quality of image captioning abilities. Our findings suggest broader significance beyond the task of generating image captions and indicate a significant shift in how artificial intelligence is applied across various domains. The addition of inclusive features shows our commitment to creating a future where technology adjusts to various user needs and helps build an artificial intelligence environment that is simpler and more welcoming.

ACKNOWLEDGMENT

The authors are thankful to AIT, Pune for providing the necessary facilities to carry out this research work.

REFERENCES

[1] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. “A Comprehensive Survey of Deep Learning for Image Captioning.” ACM Comput. Surv. 51, 6, Article 118 (November 2019), 36 pages. https://doi.org/10.1145/3295748

[2] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni and R. Cucchiara, ”From Show to Tell: A Survey on Deep Learning- Based Image Captioning,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 539-559, 1 Jan. 2023. https://doi.org/ 10.1109/TPAMI.2022.3148210

[3] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. “Deep Learning Approaches on Image Captioning: A Review.” ACM Comput. Surv. 56, 3, Article 62 (March 2024), 39 pages. https://doi.org/10.1145/3617592

[4] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj and R. K. Mishra, ”Image Captioning: A Comprehensive Survey,” 2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC),Mathura, India, 2020, pp. 325-328, https://doi.org/ 10.1109/PARC49193.2020.236619

[5] Cheng Wang, Haojin Yang, and Christoph Meinel. 2018. “Image Cap- tioning with Deep Bidirectional LSTMs and Multi-Task Learning.” ACM Trans. Multimedia Comput. Commun. Appl. 14, 2s, Article 40 (April 2018), 20 pages. https://doi.org/10.1145/3115432

[6] Bharati, P., Pramanik, A. (2020). “Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey.” In: Das, A., Nayak, J., Naik, B., Pati, S., Pelusi, D. (eds) Computational Intelligence in Pattern Recognition. Advances in Intelligent Systems and Computing, vol 999. Springer, Singapore. https://doi.org/10.1007/978-981-13-9042-5-56

[7] Shewalkar, Apeksha, Nyavanandi, Deepika and Ludwig, Simone A.. ”Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU” Journal of Artificial Intelligence and Soft Computing Research, vol.9, no.4, 2019, pp.235-245. https://doi.org/10.2478/jaiscr-2019-0006

[8] S. Degadwala, D. Vyas, H. Biswas, U. Chakraborty and S. Saha, ”Image Captioning Using Inception V3 Transfer Learning Model,” 2021 6th International Conference on Communication and Electron- ics Systems (ICCES), Coimbatre, India, 2021, pp. 1103-1108. https://doi.org/10.1109/ICCES51350.2021.9489111

[9] R. Calvin and S. Suresh,” Image Captioning using Convolutional Neural Networks and Recurrent Neural Network,” 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2021, pp. 1-4, https://doi.org/10.1109/I2CT51068.2021.9418001

[10] A. C. Hoang, D. C. Nguyen and H. L. Nguyen, ”Performance Evaluation of CNN-based Encoders for Image Captioning,” 2023 12th International Conference on Control, Automation and Information Sci- ences (ICCAIS), Hanoi, Vietnam, 2023, pp. 212-217, https://doi.org/10.1109/IC-CAIS59597.2023.10382370

[11] C. Bhatt, S. Rai, R. Chauhan, D. Dua, M. Kumar and S. Sharma, ”Deep Fusion: A CNN-

LSTM Image Caption Generator for Enhanced Visual Understanding," 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2023, pp. 1-4, https://doi.org/10.1109/CISCT57197.2023.10351389

[12] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. https://arxiv.org/abs/1409.0473

[13] T. H. Ghorpade and S. K. Shinde, "Speech Synthesis: An Empirical Analysis of Various Techniques in Text to Speech Generation," 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2023, pp. 1-6, https://doi.org/10.1109/IC-CUBEA58933.2023.10392008

[14] S. Ogun, V. Colotte and E. Vincent, "Can We Use Common Voice to Train a Multi-Speaker TTS System?," 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023, pp. 900-905, https://doi.org/10.1109/SLT54892.2023.10022766

[15] A. Nishajith, J. Nivedha, S. S. Nair and J. Mohammed Shaffi, "Smart Cap - Wearable Visual Guidance System for Blind," 2018 International Conference on Inventive Research in Computing Ap- plications (ICIRCA), Coimbatore, India, 2018, pp. 275-278, https://doi.org/10.1109/ICIRCA.2018.8597327

[16] K. B. Ram, V. B, S. P. S. Sree, C. Anilkumar, V. S. N. Reddy and B. Kodumuri, "Image Caption And Speech Generation Using LSTM And GTTS API," 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2023, pp. 992-997, https://doi.org/10.1109/ICAISS58487.2023.10250554