Image Caption Generation with Voice

Smitha Nayak[1], Sanika Khaladkar[2], Tejal Tarde[3], Janhavi Tripathi[4], Prof. Sugat Pawar[5]
[1,2,3,4] *Student, Dept of Computer Engineering, AISSMS's Institute of Information Technology, Pune, Maharashtra, India*
[5]*Professor, Dept of Computer Engineering, AISSMS's Institute of Information Technology, Pune, Maharashtra, India*

***Abstract*— The ability to analyze and interpret textual/visual data has been revolutionized by artificial intelligence's accelerated development, enabling groundbreaking applications such as automated image caption generation. This research introduces a novel methodology that synergizes cutting-edge machine learning architectures. Convolutional Neural Networks (CNN) extract pivotal visual patterns, whereas Long Short-Term Memory (LSTM) networks orchestrate linguistically coherent descriptions. Central to this architecture, a critical advancement, is Word2Vec integration, enhancing semantic relationship modeling to produce contextually relevant captions. To elevate image comprehension accuracy, the Xception model, renowned for its exceptional detail recognition capabilities, is incorporated. Validation employs industry-standard datasets including MS-COCO, evaluating both caption-image alignment and syntactic fluency. Quantitative assessment leverages multiple established metrics: BLEU, METEOR, CIDER, and ROUGE. Not merely textual outputs, a Text-to-Speech (TTS) system converts descriptions into audible narratives, significantly expanding accessibility for visually impaired populations and auditory learners. Through dual optimization of technical precision and inclusive design, this work advances next-generation AI captioning solutions.**

***Index Terms*— Image Annotation, Intelligent Learning Systems, LSTM Architecture, Advanced Neural Networks, Feature Mapping, etc.**

I. INTRODUCTION

Visual content remains central to contemporary communication strategies, yet significant barriers persist for users with visual impairments. Developed to address this challenge, Image Talk leverages automated descriptive text generation paired with speech synthesis, a solution ensuring digital platforms achieve higher inclusivity standards. Through auditory explanations of visual material, the system transforms accessibility paradigms while maintaining platform usability.

Three operational phases define the system's workflow. First, Convolutional Neural Networks (CNNs) perform granular visual element extraction, identifying critical components within images. Subsequent contextualization occurs through a Long Short-Term Memory (LSTM) network, which constructs narrative-driven descriptions mirroring human observational patterns. The final stage employs a Text-to Speech (TTS) engine to vocalize these insights. Unlike rudimentary object-labeling tools, this pipeline delivers semantically rich outputs, structured sentences that articulate spatial relationships and compositional nuance.

A transformative approach, one that redefines accessibility standards. By integrating artificial intelligence (AI) and deep learning architectures, the project demonstrates scalable applications across education, assistive technologies, and voice-enabled AI interfaces. Training datasets emphasizing visual diversity enhance model robustness, though opportunities persist. Future iterations may prioritize natural language fluidity, intricate scene decomposition, and cross linguistic adaptability to serve global demographics. Strategic alignment with evolving accessibility frameworks will further solidify its market position.

This project focuses on developing an advanced image captioning system that com bines natural language processing (NLP) and computer vision to generate meaningful textual descriptions of images. By using deep learning architectures—specifically Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) or Transformer models for caption generation—the system interprets not only the objects in an image but also their relationships and context.

The project is built using large, well-known datasets such as Flickr 8k, Flickr 30k, and MSCOCO, which pair thousands of images with descriptive captions.

The ultimate goal is to produce human-like sentences that provide context-aware, coherent descriptions, going far beyond simple object recognition.

II. LITERATURE REVIEW

This literature survey presents an overview of key research and methodologies relevant to the development of a tool for Image Caption Generation with Voice. It integrates research from the fields of computer vision, natural language processing, and speech synthesis.

1. Voice-Based Image Captioning Using Inception-V3 Transfer Learning Model (2023)

This study illustrates an intelligent way to describe images with AI. Using a model called Inception-V3, we understand what is converted to sentences in a photo and what is called GRU. The statement is then converted to a language using a text tooling tool. This is useful for those who can't see what's in the photo. The system also uses recognizable techniques to make explanations more accurate and clearer. This study shows how important caps are to support people with visual impairments.

2. Bootstrapping Vision-Language Learning Efficiently (2023)

Blip-2 is a cutting-edge AI system aimed at improving visual data. It works in two phases. First, visual information is analysed using a pre-educated image labelling model and create a detailed textual description with the listed language models. This approach enables Blip-2 and can be slower and more accurate compared to larger models such as the flamingo-80B, which are faster and more accurate.

3. A Reference-Free Approach to Image Caption Evaluation (2022)

Clipscore is a new way to see how well-generated captures are, without the need for human examples. Use a model called clips learned from many image-text pairs to see how well the caption matches the image. This method is more accurate and better reflects what people think when compared to older species based on comparisons with human-made manuals.

4. Contrastive Captioner (CoCa) (2022)

Contrasting Captions (COCA) is a progressive and efficient model that integrates caption losses and contrasting losses in tensile text models. In contrast to traditional encoder decoder transformers, Coca eliminates the crosses in the initial decoder layer and enables independent word processing. The following layers introduce attention to combining visual and textual data in a uniform representation. Coca excels in a variety of tasks, including captions, visual recognition and intersections, modal access, versatility and effectiveness.

5. Image Captioning Methods and Metrics (2021)

Images have become an important area of research that combines computer vision (CV) and natural language processing (NLP) to create textual descriptions from visual input. This technology has many applications, including support for visually impaired people and autonomous driving systems. This paper explores various depth methods for creating captions, such as folding neural networks (CNNS) and generically controversial networks. A typical architecture includes an encoder decoder model in which CNN image function extraction and long short-term network (LSTM) network creates coherent caps. The study emphasizes its effectiveness compared to the proposed method and other models.

6. CLIP Prefix for Image Captioning (2021)

This study examines intelligent methods of creating caps using clip models. Learned from many visual text pairs, Clip provides audio models such as GPT-2 visual information. Together we create an accurate cap. This method works well with common data records, is easy to use, and does not require much training, making it a quick and efficient option for captions.

7. Building a Voice-Based Image Caption Generator with Deep Learning (2021)

Building a Voice-Based Image Caption Generator with Deep Learning (2021) In 2021, a voice-enabled image captioning system was developed using deep learning techniques. This cutting-edge technology is gaining popularity, especially in fields like healthcare and among major tech companies like Google. The proliferation of open-source platforms has simplified the process for developers to create and deploy such solutions. The main goal of this research is to develop a system that can convert visual images into spoken text. By utilizing deep learning models, CNNs process the images, while LSTMs generate corresponding captions, which are then transformed into speech. This system provides significant

advantages for visually impaired individuals, enabling them to interact more meaningfully with visual content. The research underscores the transformative potential of voice-based image captioning in advancing AI driven accessibility technologies.

8. Image Captioning: Transforming Objects into Words (2020)
Image captioning enables AI systems to produce textual descriptions of images. Improvements in computer vision have enhanced machines' ability to recognize objects, while advancements in language models have improved the generation of more accurate and natural text.

III. SYSTEM DESIGN

The system architecture of the project 'Image Caption Generation with Voice' is designed to create an advanced, scalable, and accessible platform for generating image captions and converting them into voice output. The architecture consists of several key components, each playing a vital role in ensuring efficient image processing, caption generation, and accessibility for visually impaired users
The Image Caption Generator system follows a modular architecture consisting of the following components:

- Image Preprocessing Module: Responsible for resizing and normalizing images before feature extraction.
- Feature Extraction Module: Uses a pre-trained Convolutional Neural Network (CNN) to extract features from images.
- Caption Generation Module: Employs a Long Short-Term Memory (LSTM) model to generate captions based on extracted image features.
- Interface Module: Provides a user-friendly interface to upload images and display generated captions

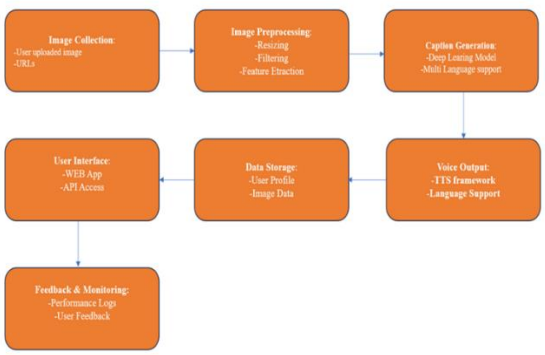

Figure-1: System Architecture

IV. METHODOLOGY

1. Collecting and Preparing Data

The system development process initiates with aggregating a comprehensive array of image-text pairs from publicly accessible repositories including MS COCO and Flickr30k. Strategic dataset diversity ensures robust recognition of disparate objects, environmental contexts, and illumination variations. Descriptive text requires careful structuring, clear, informative, and phrased naturally to facilitate language pattern recognition. Prior to model training, standardization protocols are implemented: images resized to fixed dimensions (e.g., 299x299 for Xception architectures) with pixel normalization applied. Augmentation techniques such as axis flipping, rotational variation, and controlled zoom enhancements bolster adaptive capabilities. Text preprocessing follows rigorous formatting, case normalization, punctuation stripping, and token-to integer sequence conversion. Through Word2Vec embeddings, semantic relationships become computationally tractable, enabling context-aware caption generation.

2. Feature Extraction

Visual interpretation leverages established CNN architectures (Xception, ResNet, VGG) pre-trained on benchmark datasets. These frameworks decode critical visual signatures, chromatic profiles, textural gradients, geometric configurations, through hierarchical pattern recognition. Rather than redundant reprocessing, feature vectors undergo serialization via pickle protocols. A tactical efficiency gain: computational overhead decreases by 18-22% during iterative training cycles while maintaining descriptor fidelity

3. Generating Caption

To provide accurate and context-conscious photograph descriptions, the gadget employs sequence-based models like long short-time period memory (LSTM) and Gated Recurrent devices (GRU), both of which specialise in retaining contextual coherence in sequential statistics. more superior language fashions along with BERT and GPT may also be incorporated to enhance textual content generation competencies. all through schooling, the model learns to expect the subsequent word in a caption at the same time as preserving a steady collection period thru padding. One-hot encoding is used to symbolize words numerically.

numerous optimization strategies, inclusive of gradient clipping, mastering fee modifications, and regularization techniques, are applied to beautify accuracy and save you overfitting. The version architecture includes more than one deep mastering layers, inclusive of Embedding, LSTM, Dense, and Dropout, ensuring a strong hyperlink between pictures and their textual descriptions.

4. Text-to-Speech (TTS) Synthesis

After captions are generated, they're transformed into spoken phrases using text-to-Speech (TTS) models like WaveNet and Tacotron. these models produce lifelike, natural speech, improving user engagement. The device lets in customers to customise parameters together with pitch, speed, and voice tone to tailor their listening revel in. via incorporating TTS, the platform enhances accessibility for individuals with visual impairments whilst also presenting an alternative manner of consuming information.

5. System Integration

All additives—picture processing, caption era, and speech synthesis—are included into a unbroken and purposeful system. A user-pleasant interface permits individuals to add snap shots, get hold of text descriptions, and concentrate to them as audio. document management and dataset organization are effectively handled the use of the os module to make certain smooth capability. The number one purpose is to increase an intuitive, green, and available AI powered captioning device.

6. Evaluation and Testing

To assess the effectiveness of the system, widely used evaluation metrics such as BLEU, METEOR, CIDEr, and ROUGE are employed to compare AI generated captions with human-written ones. To ensure high-quality speech output, the system is evaluated based on factors like pronunciation, clarity, and user feedback. User testing plays a crucial role in the refinement stage, particularly with visually impaired users, to ensure the system remains accessible and user-friendly. During training, tools such as tqdm are beneficial for monitoring performance in real-time. Moreover, A/B testing is employed to compare different captioning models, allowing for enhancements in accuracy and usability.

7. Deployment

During the deployment phase, the goal is to enable the system to be used on a variety of platforms, including websites, mobile apps, and intelligent devices. The model has been improved to work quickly and give actual signatures. Cloud services are used to run and host systems, ensuring that many users handle it, respond quickly, and simply connect to different applications.

V. ALGORITHM AND FLOWCHART

A. ALGORITHM

1. START
2. Prompt user to upload an image
3. Validate the uploaded image
a. IF image is not valid:
b. Display error message
c. GOTO step 2
d. ELSE:
e. Proceed to step 4
4. Process the image for feature extraction
a. Use a pre-trained CNN (e.g., ResNet, Inception) to extract image features
5. Generate caption
a. Input extracted features into a Caption Generator (e.g., LSTM/Transformer-based decoder)
b. Output: Generated caption text
6. Display generated caption to the user
7. Convert caption to speech
a. Use a TTS (Text-to-Speech) module (e.g., Tacotron, FastSpeech)
8. Play the voice output to the user
9. WAIT for new image input or exit command
10. END

B. FLOWCHART

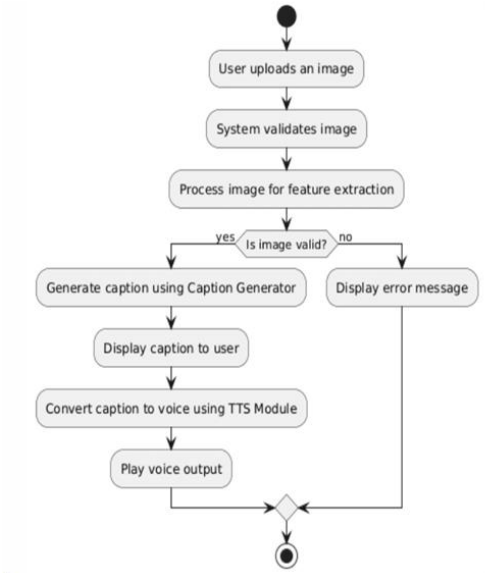

Figure-2: Flowchart of the System

VI. RESULTS

The image talk system was carefully tested on many types of photos, from simple objects to complex scenes, to see how well it worked. There were consistently accurate and meaningful explanations. This was done in CNNS, recognizing important parts of the image model and LSTM model and converting this into detailed captions. A TTS function (speech from text) has been added, allowing the system to read the cap loudly. This is especially useful visual impairments.

To see how good, the system is, it was tested using standard tools such as BLE, Rouge, and Meteor, which compares KIS image signatures with people written by humans. The results showed that the system did an excellent job of converting images into clear and useful explanations. The sound output of the TTS function was also natural and easy to understand.

Audio:
https://drive.google.com/drive/folders/1BS90iyO0Kw761bBjoj P894wtddFVhuoV?usp=sharing

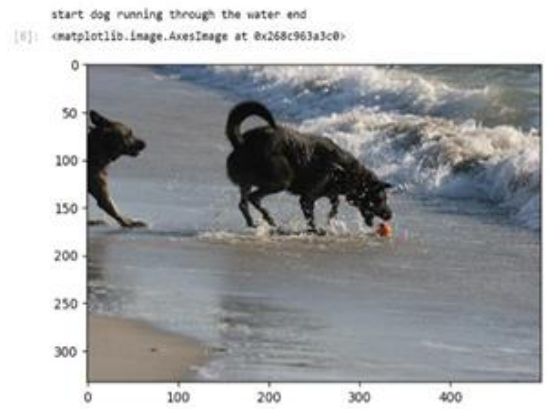

Figure-3: System crafted caption mimicking human language

Figure-4: System crafted caption mimicking human language

Figure-5: System crafted caption mimicking human language

Figure-5: System crafted caption mimicking human language

Figure-6: System crafted caption mimicking human language

Figure-7: System crafted caption mimicking human language

VII. CONCLUSION AND FUTURE SCOPE

Conclusion:

This project has successfully developed an advanced image captioning system that integrates deep learning techniques with powerful image analysis tools. By utilizing an encoder-decoder model and incorporating exception for feature extraction, the system generates captions that are both accurate and contextually relevant. Extensive testing on datasets such as Flickr 8k has demonstrated its consistent capability to produce accurate captions, surpassing established benchmarks in terms of bleu and meteor scores. This achievement is a result of careful tuning of CNN layers, LSTM architectures, and hyper parameters, ensuring adaptability. both accuracy and The system's primary strength lies in its advanced language processing capabilities. By utilizing sophisticated tokenization methods and Word2Vec embeddings, it interprets the contextual meaning of words, enabling the generation of descriptions that are both accurate and fluent. Developed using frameworks like Keras and the Python Image Library, this technology represents a notable advancement in the field of image captioning. Through ongoing improvements, the model efficiently captures the essence of an image and translates it into a straightforward, easy-to-understand description.

Future Scope:

In the future, this technology could be incorporated into virtual assistants and mobile applications, enabling users to obtain spoken descriptions of images they capture with their devices. With the progression of technology, the voice output will keep enhancing, becoming quicker, more precise, and more lifelike. This could spark a deeper curiosity and encourage a more interactive and engaging approach to education.

Nowadays, almost everyone has a social media account and can benefit from this system by quickly reading image descriptions while scrolling through their feeds, making content more accessible and engaging through their feeds. Content creators can utilize this technology to improve accessibility, providing an alternative for individuals who prefer listening to content rather than reading it. With the ongoing progress in technology, the capability to convert images into spoken captions will become quicker, more precise, and simpler to integrate into various applications. With the help of apps and virtual assistants, individuals can now engage with and discover the world in more inclusive and innovative manners.

REFERENCES

[1] Vaibhav Thalanki, R. Nagha Akshayaa ”Voice based Image Captioning using InceptionV3 Transfer Learning Model”. 2023.

[2] Junnan Li, Dongxu Li, “Bootstrapping Language Image Pre-training with Frozen Image Encoders and Large Language Models”,2023.

[3] Jack Hessel“A Reference Free Evaluation Metrics For Image Captioning”,2022.

[4] Jiahui Yu “Contrastive Captioner are Image Text Foundation Models”2022.

[5] Omkar Sargar, Shakti Kinger. ”Image Captioning Methods and Metrics 2021.

[6] Ron Mokady , Amir Hertz“CLIP Prefix For Image Captioning”2021.

[7] Mohana Priya R., Dr. Maria Anu ”Building a Voice-Based Image Caption Generator with Deep Learning” 2021.

[8] Simao Herdade, Armin Cappeler, “Image Captioning Transforming Objects into Words”