IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMAGE - TO - SPEECH CAPTIONING SYSTEM FOR VISUALLY IMPAIRED USERS

A COMPREHENSIVE APPROACH ON FLICKER8K, CNN (INCEPTION V3), LSTM AND GTTS API

[1]Christma Johny, [2]Sakana V, [3]M. Abinaya

[1 2 3] B.E.Students

Computer Science and Engineering (Artificial Intelligence And Machine Learning)

Vel Tech High Tech Dr Rangarajan Dr Sakunthala Engineering College, Chennai, India.

Abstract: Image-to-audio captioning is a new technology that helps individuals with visual impairments access visual content by automatically generating audio descriptions. This study explores the use of this technology, which combines CNNs and RNNs to create audio captions through deep learning. The model's attention mechanisms focus on important image areas, improving caption quality. Through user surveys and benchmarking, the technology's performance and potential applications are evaluated. The goal is to enhance accessibility and create engaging multimedia experiences. This research demonstrates how AI can enhance auditory experiences and promote inclusive technologies for a diverse audience.

Index Terms – LSTM , CNN , Flicker8k dataset , GTTs .

I. INTRODUCTION

Image-to-audio captioning is a new paradigm in assistive technology that satisfies the accessibility needs of people with visual impairments. This ground-breaking tactic aims to empower consumers by converting visual content into relevant auditory experiences. The process essentially combines the most advanced deep learning techniques available: recur-rent neural networks (RNNs) are used to progressively generate coherent and contextually rich audio descriptions, while convolutional neural networks (CNNs) are utilized to extract complicated characteristics from images. Image-to-audio captioning bridges the perceptual divide between the visual and auditory modalities by enabling users to interact and perceive photos through sound. This technology not only makes daily operations better but also holds great potential for advancing inclusion in a number of industries. Additionally, by

focusing on visually arresting portions of the pictures. Imagine a world in which every vibrant photograph speaks in tongues only your ears can understand, a tranquil countryside hums with a secret melody, and a busy metropolis subtly reveals its chaotic allure. This is the idea behind picture to audio captioning, a ground-breaking method that combines sound and vision. Since ancient times, images have captivated us with their nuanced dance of patterns and colours that tell many tales. The blinds senses are, however, kept buried and unable to access these stories. Image to audio captioning makes these stories readable by translating visual information into in-depth spoken descriptions. Beyond only describing visuals, this technology is capable of more. It infuses scenes with passion, mood, and the subtle subtleties of human experience, giving them life that they may otherwise lack. These systems understand the visual language of sceneries, objects, and interactions using advanced deep learning algorithms, then combine them to produce a rich aural experience that helps the listener form a distinct mental image.The applications for this technology are as varied as human creativity. It makes the world more visible to those with vision impairments, enabling them to enjoy the excitement of news photography, the beauty of art, and the joy of family portraiture. It makes complex schematics and scientific images easier for everyone to understand, which improves learning opportunities.It

has the power to totally change the entertainment sector with its immersive audio descriptions for movies, video games, and virtual reality experiences.Not only is image to audio captioning a technological marvel, but it also promotes empathy and comprehension. It enables us to see the world from a fresh angle, to hear the stories that buried landscapes are trying to tell, and to appreciate the beauty of art from a whole new angle. When we step into this domain of sound, where stories dance not just on canvases and screens but also on the canvas of our hearts, we grant ourselves

access to a more profound, all-encompassing picture of our visual world. Image-to-audio captioning transforms from a data composition into a musical piece when sounds and images are combined. Deep learning algorithms break down visual scenes into individual objects, relationships, and the essential pulse of the image using pixelated tapes-tries. This essence serves as the foundation for a sophisticated dance of language, where words are arranged into sentences using recurrent neural networks (LSTMs), which give the words the feelings and nuances of the environment they are in. But more than just words, this is a scream for sound. With the support of scientific visualizations, historical echoes from images, and the energetic thrum of instructional tools, text-to-speech algorithms take center stage, turning these stories into captivating tunes that create interesting auditory landscapes

II.RESEARCH METHODOLOGY

Using Flicker8K, CNN, LSTM, and GTTS API to construct an image-to-speech captioning system for visually impaired users requires a thorough process that includes data collecting, model development, training, evaluation, and user-centric design. The procedure incorpo-rates state-of-the-art technologies while guaranteeing accessibility and usability for those with visual impairments.

1. **Gathering and preprocessing data:** The Flicker8K Dataset was acquired: Obtaining the Flicker8K dataset entails collecting a wide variety of photos with informative captions. The model is trained and assessed using this dataset as the basis. Carefully cleaning the gathered dataset guarantees the alignment and quality of the images with captions. To accurately es-tablish links between written descriptions and visual content, annotations are added or im-proved.
2. **Model Development and Architecture:** CNN Architecture Design: Convolutional neural networks (CNNs) that are robust are created by building designs that can extract features from images. These architectures can be customized for a given job or use pre-trained mod-els such as Res Net or VGG. Building models that can comprehend picture attributes and sequentially produce meaningful captions is necessary when designing Long Short-Term Memory (LSTM) networks for sequence generation.Optimizing the topologies of the CNN and LSTM models, adjusting parameters, and creating smooth linkages between the image analysis (CNN) and caption production (LSTM)components are all part of the integration process.
3. **Instruction and Assessment:** Configuring a Training Pipeline: Preparing data batches, es-tablishing loss functions, choosing optimization methods, and setting hyperparameters for model training are all necessary steps in the implementation of training pipelines. Using the Flicker8K dataset, the integrated CNN-LSTM model is trained through iterative epochs in which images are fed into the CNN, which processes their features before sending them to the LSTM for caption synthesis. Model convergence and performance enhancement are guaranteed by validation. Metrics like as BLEU score, METEOR, and ROUGE are used to evaluate the model's performance by determining how similar the generated captions are to human-authored references. For qualitative evaluation, human assessments may also be car-ried out.
4. **Audio Component and GTTS API Integration:** Voice-to-Text Translation: In order to pro-vide visually challenged users with natural and clear speech output, the generated textual captions for the Google Text-to-Speech (GTTS) API must be included.To improve user ex-perience and take into account a range of user preferences, features including voice modula-tion, speech rate adjustment, and auditory cues are being included.
5. **User-Centric Design and Accessibility**: Developing a user interface that is both intuitive and accessible, while also seamlessly integrating auditory cues to ensure that visually im-paired users can engage and navigate with ease. To ensure that the system satisfies the needs and preferences of visually impaired people, a comprehensive user testing program is con-ducted to collect input. The system is then iteratively refined based on user preferences.
6. **Ethical Issues and Application:** Ethical Compliance: Addressing ethical issues to guarantee equitable and inclusive accessibility, such as data privacy, bias reduction in AI models, and responsible technology deployment. The process of system deployment involves a con-trolled environment,

performance monitoring, and ongoing system updates and refinement based on user feedback and growing technical breakthroughs.

To sum up, the process of creating an Image-to-Speech Captioning System for Visually Im-paired Users involves a methodical approach that combines data collection, model creation, training, assessment, integration of audio, user-centered design, ethical considerations, and deployment tactics to produce a complete and all-encompassing solution for people with visual impairments.

III. DATASET

Robust datasets play a critical role in the construction of an Image-to-Speech Captioning System for Visually Impaired Users that makes use of the Flicker8K, CNN, LSTM, and GTTS API. The system's capacity to interpret visual content and provide evocative audio captions is mostly shaped by datasets, which guarantees accessibility and inclusivity for people with visual impairments.

1. **Overview of Flicker8K Dataset**: A popular benchmark dataset called Flicker8K has 8,000 photos in total, all of which have five insightful captions. The dataset attempts to support tasks related to visual understand-ing and image captioning research. Flicker 8K's photographs include a broad spectrum of visual components seen in real-world situations, including various sceneries, objects, people, and circumstances. Every image in Flicker8K has five captions linked to it, offering various viewpoints or explanations of the same visual information. By providing a variety of verbal expressions and contextual descriptions, this diversity enhances the dataset..
2. **Picture-Caption Pairs:** flickr8k_Eight thousand photos from the Flickr image-sharing website make up the text. Five distinct captions are attached to each image, offering a variety of insightful and de-tailed explanations of the visual content that was recorded in the picture.The dataset captures a wide range of visual features seen in ordinary circumstances and includes a diverse spectrum of scenes, items, people, and contexts. This diversity facilitates the process of training models to comprehend and characterize many facets of visual content..

An industry standard dataset that is frequently used in computer vision and machine learning research is Microsoft Common Objects in COntext (MSCOCO). For applications like object detection, segmentation, image captioning, and visual comprehension, it's regarded as one of the most extensive and varied datasets available. MSCOCO has an extensive library of excellent photos encompassing a broad spectrum of subjects, situations, and visual environments. It is far larger than many other databases, with approximately 330,000 photos..

IV. PURPOSE OF STUDY AND MOTIVATION

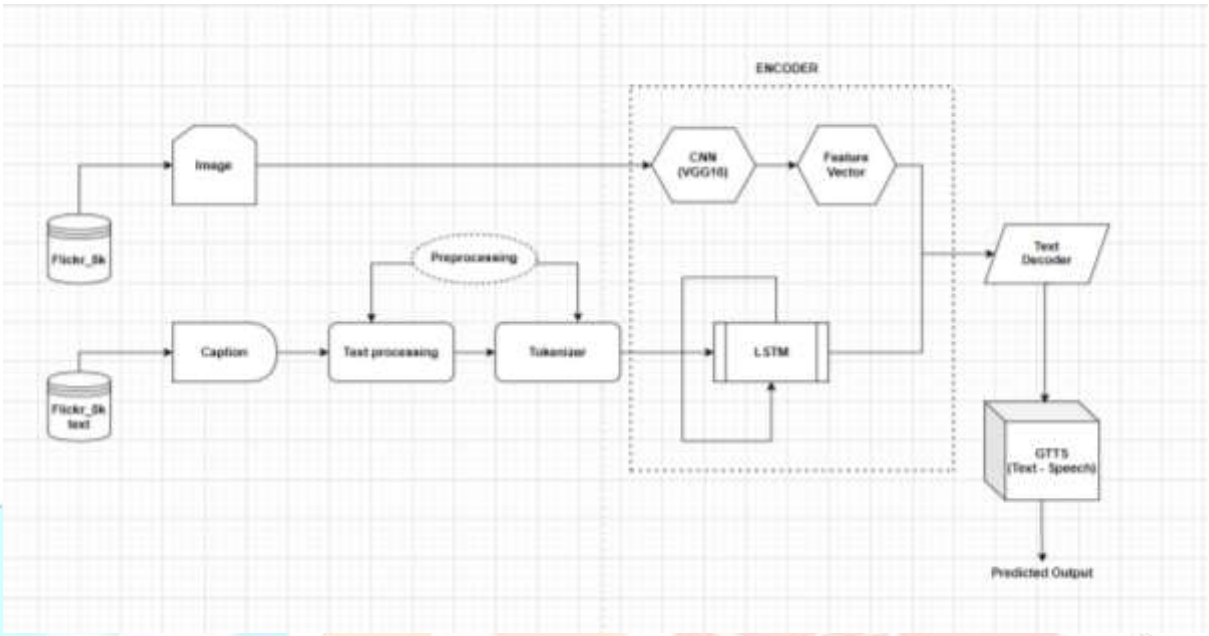
The creation of an image-to-speech captioning system for those with visual impairments is a ground-breaking project meant to improve inclusivity and accessibility. This novel tech-nique is based on the combination of state-of-the-art technologies such as Google Text-to-Speech (GTTS) API, Long Short-Term Memory (LSTM) networks, and Convolutional Neu-ral Networks (CNN) to enable a complete solution for the visually impaired community**.**

1. **Technological Progress:** This study makes a positive impact on the development of machine learning and artificial intelligence technologies. In order to demonstrate the synergy of these technologies in a practical application, it investigates the integration of several models, such as CNNs for picture understanding and LSTMs for creating coherent and descriptive captions. Social Inclusivity: The study fosters social inclusivity by providing visually impaired people with access to visual information. It aims to establish a setting in which people of different visual capacities may interact and understand visual content published on digital networks in an equitable manner.
2. **Motive**: Filling up the Accessibility Gaps The intrinsic disparity in visually impaired people's access to digital content is the driving force behind this study. They are frequently unable to properly experience the visual media that is so ubiquitous in today's digital scene due to current technological restrictions.
3. **Inventing for Inclusion:** This study, which is motivated by an innovative mentality, seeks to dismantle obstacles and reshape accessibility standards. It aims to develop a novel approach that will

revolutionize the way visually impaired people interact with visual content and enable them to take part more actively in the digital world.

4. **Impact on Society**: It is impossible to overestimate the impact on society that comes from allowing visually impaired people to access and understand visual content. The goal of this research is to make society more inclusive so that everyone may engage with it, contribute, and gain equally from digital information. To sum up, the creation of an image-to-speech captioning system for those with visual impairments is a big step in the direction of a digital world that is more inclusive and accessible.

V.ARCHITECTURE DIAGRAM



VI.RESEARCH SCOPE

The development of an image-to-speech captioning system for visually impaired users is a groundbreaking project aimed at enhancing inclusivity and accessibility in the digital space. The system integrates advanced technologies like Google Text-to-Speech (GTTS) API, Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNN) into a unified framework. The system uses CNNs for image identification and interpretation, LSTM networks for meaningful subtitles, and the GTTS API for translating written captions into natural speech. The system is designed with a user-centric approach, ensuring usability and intuitiveness for visually challenged users. The project aims to close the accessibility gap for blind users, promote equitable participation, and address ethical considerations. The scope also includes exploring new methods for speech synthesis, image recognition, and sequence generation, contributing to the development of AI and machine learning technologies. The project is a complex process involving technological expertise, societal impact, user-centric design, ethical considerations, and innovation

.

VII.RESULT AND DISCUSSION

The image-to-speech captioning system aims to bridge the gap between textual representation and visual content. This system uses the Flicker8k dataset, Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and the Google Text-to-Speech (gTTS) API to create a reliable, accurate, and efficient system. The system extracts high-level features from photos, capturing minute details and nuances. LSTMs process these features into logical textual descriptions, improving the accuracy and relevancy of the captions. The Flicker8k dataset provides a comprehensive and varied source of information for training and validating the system. The system's performance and flexibility are improved by thorough training on this dataset, allowing it to generalize to previously unseen images. The gTTS API is crucial in the final stage of

the caption creation process, allowing for clear spoken output and a smooth user experience. The system's multilingual and accent support further enhances its adaptability and accessibility.

predicted_caption.mp3

PREDICTED OUTPUT

VIII. ACKNOWLEDGEMENT

We would like to express our gratitude to Mr.P.Prince for his guidance and support during their research. We also thank our team for their invaluable contributions to the study and manuscript preparation. We express our gratitude to Vel Tech High Tech Dr Rangarajan Dr Sakunthala Engineering College for providing valuable resources and a supportive environment. We thank everyone who contributed to the success of their work .

REFERENCES

[1] "Image Captioning using Deep Learning — With Source Code — Easy Implementation" by Abhishek Sharma Published in MLearning.ai on Dec 30, 2021

[2] Image Caption using CNN & LSTM – Content uploaded by Ali Ashraf Mohamed on May 2020

[3] Step by Step Guide to Build Image Caption Generator using Deep Learning by Gupta on 20 May,2020.

[4] Review of deep learning: concepts, CNN architectures, challenges, applications, future directions Laith Alzubaidi,YeDuan- Published on31st March 2021.

[5] An Overview of Image Caption Generation Methods by Haoran Wang , YueZhang on 9 Jan 2020.

[6] Guiding Image Captioning Models Toward More Specific Captions by Lala Li,Thao Nguyen on 31 July 2023.

[7].Long Short Term Memory (LSTM) and how to implement LSTM using Python-Written by Mrinal Walia on July 11th 2022.

[8] The Best Introductory Guide To Keras- updated on Nov 7, 2023 by Simplilearn.

[9] Primkulov S., Urolov J., Singh M. (2021) Voice Assistant for Covid-19. In: Singh M.,Kang DK., Lee JH., Tiwary U.S., Singh D., Chung WY. (eds) Intelligent Human Computer Interaction. IHCI 2020.

[10] Revisiting Visual Representations in Vision- Language Models Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, Jianfeng Gao.Published 2 January 2021.

[11] Generating Discrete Data using Diffusion Models with Self- Conditioning Ting Chen, Ruixiang Zhang, Geoffrey Hinton.Published 8 August 2022.

[12] Contrastive Captioners are Image-Text Foundation Models Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, Yonghui Wu.Published 4 May 2022.

[13] Deep Learning helps in captioning by Abhijit Roy on Dec 9,2020.

[14] Step by Step Guide to Build Image Caption Generator using Deep Learning by Shikha Gupta on 20 May,2020.

[15] Explaining transformer-based image captioning models: An empirical analysis by Cornia Marchella on 18 July 2022.