MULTILINGUAL IMAGE CAPTIONING WITH VOICE ASSISTANCE

A Project Report submitted by

Krishna Kumari Ravuri

in partial fulfillment of the requirements for the award of the degree of

M.Tech.

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology Jodhpur
Department of Mathematics and Computational Science
November 2023

Declaration

I hereby declare that the work presented in this Project Report titled MULTILINGUAL IMAGE CAPTIONING WITH VOICE ASSISTANCE submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of M.Tech., is a bonafide record of the research work carried out under the supervision of Dr. Sukhendu Ghosh of the respective Supervisor. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

R.K.Kumari

Signature

KRISHNA KUMARI RAVURI

M22AI567

Certificate

This is to certify that the Project Report titled MULTILINGUAL IMAGE CAPTIONING WITH VOICE ASSISTANCE, submitted by Krishna Kumari Ravuri (M22AI567) to the Indian Institute of Technology Jodhpur for the award of the degree of M.Tech., is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature
Dr. Sukhendu Ghosh

Abstract

This project endeavour to address the significant educational and employment challenges faced by the visually impaired population in India due to limited access to Braille and expensive assistive technologies. With less than 10 percent of blind individuals having access to Braille education, a critical literacy gap exists. To mitigate this, our proposal outlines the development of an affordable and accessible solution utilizing computer vision, object recognition, and speech synthesis technologies. The envisioned model aims to convert text from images into machine-readable format and subsequently employ speech synthesis to relay real-time information about the user's surroundings. A key focus is the incorporation of regional languages, given India's linguistic diversity, ensuring the model's capability to read aloud text in multiple languages. This comprehensive approach seeks to enhance the quality of life and open up new opportunities for the visually impaired community in India.

Keywords: Computer vision, Object recognition, Speech synthesis, Multilingual text recognition, Machine learning , convolutional neural network (CNN),recurrent neural network (RNN)

1. INTRODUCTION

In the vast tapestry of India's societal landscape, the visually impaired encounter substantial challenges in the realms of education and employment, primarily stemming from a glaring lack of accessible resources. The deficiency in Braille education, coupled with the exorbitant costs associated with assistive technologies, has resulted in a poignant literacy gap, leaving less than 10 percent of blind individuals with the means to learn Braille.

One distinctive aspect of our project lies in its commitment to linguistic diversity. India, with its multitude of languages and dialects, presents a unique challenge and opportunity. Recognizing this, our model is meticulously designed to accommodate and proficiently handle multiple regional languages. Beyond the technological intricacies, our emphasis on linguistic inclusivity is a conscious effort to ensure that the model can seamlessly convert and articulate text in diverse languages, resonating with the linguistic preferences of the user.

This linguistic adaptability is not merely a feature but a foundational principle that aims to foster inclusivity and usability across India's diverse linguistic spectrum. This adaptability resonates with the linguistic preferences of users, transcending technological intricacies to promote inclusivity across India's expansive linguistic spectrum.

At the core of our solution resides a sophisticated technological framework seamlessly integrating multiple components. Employing a pre-trained convolutional neural network (CNN), specifically InceptionV3, empowers our model with robust image feature extraction. Complementing this is an attention mechanism that enhances the model's focus on pertinent aspects of the input image during caption generation. The recurrent neural network (RNN)-based decoder further refines the process, generating captions based on the extracted features to construct a coherent and contextually relevant description of the user's environment.

This holistic technological approach not only addresses immediate challenges faced by the visually impaired but positions our solution as a scalable and adaptive assistive tool for the future. Through this fusion of technological prowess and inclusivity, our project endeavours to propel meaningful advancements in the educational and employment landscape for the visually impaired community in India.

2. BACKGROUND

impaired population in India, particularly in the realms of education and employment. Existing barriers, such as limited access to Braille resources and the high costs associated with assistive technologies, contribute to a pronounced literacy gap among the visually impaired, with fewer than 10 percent having the opportunity to learn Braille. Recognizing the critical need for innovative solutions, this project seeks to leverage advancements in computer vision, object recognition, and speech synthesis technologies to create a transformative tool.

The motivation to address these challenges is rooted in a broader commitment to inclusivity and empowerment. The dearth of accessible resources and the lack of affordable assistive technologies not only limit educational opportunities but also impede access to meaningful employment for the visually impaired. This project aims to fill this void by developing a sophisticated yet accessible model that harnesses the power of technology to enhance the quality of life for the visually impaired community in India.

Furthermore, the linguistic diversity of India adds a layer of complexity to this challenge. The project acknowledges and seeks to address the rich tapestry of languages and dialects spoken across the country. By designing the model to read aloud text in multiple regional languages, the project aims to ensure that the solution is not only technologically robust but also culturally and linguistically sensitive.

In the broader context of technological advancements and social responsibility, this project envisions contributing to the creation of a more inclusive and equitable society. By bridging the educational and employment gaps for the visually impaired in India, the project aligns with a commitment to leveraging technology for social good and empowering marginalized communities. Through a combination of cutting-edge technology and a deep understanding of societal challenges, the project endeavour to make a meaningful impact on the lives of the visually impaired in India.

3. LITERATURE SURVEY

1. Global Overview:
The World Health Organization's "World Report on Vision" provides a comprehensive global perspective on visual impairment, offering insights into the challenges faced by visually impaired individuals worldwide. This report is instrumental in understanding the broader context of visual impairment and initiatives on a global scale.

2. Assistive Technologies Worldwide:
The RNIB's "Technology for Life: Assistive Technology Report" is a valuable resource for exploring successful global models of assistive technologies. It highlights innovations and best practices that have proven effective in enhancing the lives of visually impaired individuals, serving as a benchmark for the proposed project.

3. Indian Perspective on Assistive Technologies:
The academic paper by George and Mathew (2018) delves into "Assistive Technologies in Education for Visually Impaired: An Indian Perspective." This work specifically addresses the Indian context, providing insights into existing technologies, challenges faced, and potential areas for improvement in education.

4. Saksham's Technology Initiatives:
"Saksham: Technology for the Blind" offers on-the-ground insights into technology initiatives for the blind in India. The organization's website may provide real-world case studies and user testimonials, contributing to a better understanding of the local landscape and potential project applications.

5. Multilingual Text Recognition Systems:
Singh and Yadav's (2020) academic review on "Multilingual Text Recognition Systems" explores the landscape of systems crucial for the project's goal of regional language adaptability. It provides insights into challenges and solutions in recognizing and vocalizing text in diverse languages.

6. ICT Accessibility in Indian Languages:
The resource from the Centre for Internet and Society (2018) titled "Accessibility of ICTs in Indian Languages" offers an in-depth exploration of the accessibility of Information and Communication Technologies in Indian languages. It sheds light on challenges related to language diversity and the development of inclusive technologies.

7. Ethical Considerations in Assistive Technologies:
Foley and Mathews (2019) discuss "Ethical Considerations in the Use of Assistive Technology for People with Disabilities." This academic paper guides the project in ensuring ethical principles, particularly regarding privacy and dignity, are upheld in the development of the proposed solution.

8. Be My Eyes Case Study:
"Be My Eyes," a real-world implementation providing remote assistance to visually impaired individuals, serves as a relevant case study. This application showcases the potential of technology to connect and improve the lives of visually impaired individuals.

9. Language Processing for Indian Languages:
Ghosal and Hazarika's (2019) academic review on "Language Processing for Indian Languages" provides insights into the challenges and advancements in linguistic diversity. It informs the project's approach to handling multiple regional languages.

10. Microsoft Translator for Education:
The Microsoft Translator for Education offers a practical tool for language adaptability. This resource aligns with the project's goal of enhancing educational opportunities for the visually impaired through translation services in multiple languages.

This condensed literature survey synthesizes a range of global and local perspectives, technology initiatives, case studies, and academic reviews that collectively inform the proposed project on developing assistive technology for the visually impaired in India.

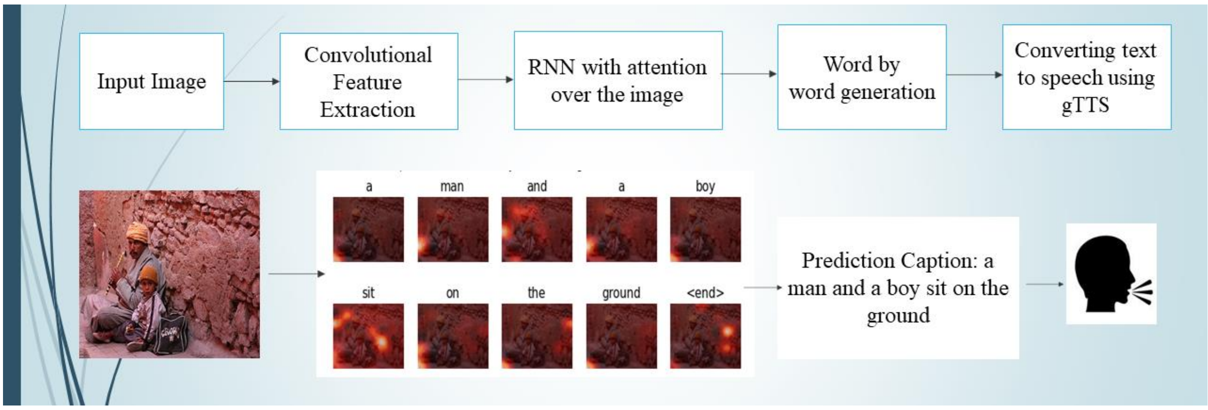
Input Image
Convolutional Feature Extraction
RNN with attention over the image
Word by word generation
Converting text to speech using gTTS
a
man
and
a
boy
sit
on
the
ground
<end>
Prediction Caption: a man and a boy sit on the ground

Fig.2. Proposed System Architecture

4. PROBLEM DEFINITION

The visually impaired population in India encounters formidable challenges in accessing education and employment opportunities due to a critical lack of accessible technologies. Less than 10 percent of blind individuals in India have the opportunity to learn Braille, resulting in a substantial literacy gap. This stark reality limits their access to information and inhibits their ability to navigate their surroundings independently. The absence of affordable and inclusive assistive technologies exacerbates the difficulties faced by the visually impaired, hindering their quality of life and impeding their integration into educational and professional spheres. To address these pressing issues, there is a crucial need for an innovative solution that combines computer vision, object recognition, and speech synthesis technologies. This project aims to develop a model capable of converting visual information into audible text, providing real-time environmental insights and thereby enhancing the overall quality of life and opportunities for the visually impaired in India. The key challenge lies in ensuring the model's adaptability to India's linguistic diversity, making it proficient in converting images to text and vocalizing content in multiple regional languages. By tackling these challenges, the project endeavours to bridge the existing gaps, empowering the visually impaired community to navigate their surroundings more effectively and participate more fully in educational and professional endeavours.

5. OBJECTIVE

Design and implement an affordable and accessible assistive technology solution leveraging computer vision, object recognition, and speech synthesis technologies. The primary objective is to create a tool that overcomes the financial barriers faced by the visually impaired in accessing advanced assistive technologies.

Enable the model to process visual information in real-time, providing the visually impaired with immediate insights into their surroundings. This includes reading text aloud, identifying objects, and offering contextual information to enhance their overall environmental awareness.

Ensure the model's proficiency in converting images to text and reading out information in multiple regional languages. The objective is to make the technology inclusive and culturally relevant for the diverse linguistic landscape of India, enhancing its usability and impact.

Facilitate better access to education for the visually impaired by providing real-time audio feedback on written content. The model should be capable of reading textbooks, study materials, and other educational resources aloud, thereby promoting inclusive learning environments.

Support the visually impaired in their pursuit of employment by providing real-time information about their work environment. This includes reading written documents, identifying objects, and offering audio cues to enhance their workplace navigation and productivity.

Prioritize a user-centric design, focusing on human-computer interaction principles to create an intuitive and user-friendly experience. The objective is to empower visually impaired users with a technology solution that is easy to navigate and seamlessly integrates into their daily lives.

6. METHODOLOGY

The methodology for the proposed image captioning system encompasses a multifaceted approach. Initially, raw input data, comprising images and associated captions, undergoes meticulous pre-processing to ensure uniformity and quality. Subsequently, a pre-trained Convolutional Neural Network (CNN) serves as an encoder, extracting and encoding salient features from the images into a higher-dimensional vector space. The encoded features act as a meaningful input for the subsequent decoding process. The decoding process involves an LSTM-based Recurrent Neural Network (RNN), functioning as the decoder, which translates the encoded features into coherent natural language descriptions. To enhance the overall performance and contextual relevance of the generated captions, an attention mechanism is employed, allowing the model to selectively focus on distinct regions of the input image during the caption generation process. Importantly, the methodology extends beyond caption generation by integrating Google Translator for caption translation into multiple languages. This post-processing step ensures that the generated captions are not only contextually rich but also accessible in a multitude of languages, addressing the linguistic diversity imperative for widespread usability. The final step involves utilizing beam search to refine and select the most probable captions. This comprehensive methodology amalgamates state-of-the-art techniques in computer vision, natural language processing, and translation services to create a versatile and inclusive image captioning system.

7. REFERENCES

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Zemel, R. (2015). Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention.

2. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator.

3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.

4. Lu, J., Xiong, C., Parikh, D., & Socher, R. (2015). Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning.

5. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context.