Image Caption Bot for Assistive Vision

Prof.Anandkumar Birajdar*, Pranav Patil, Kunal Patil, Ketan Sarode, Omkar Rajhans

Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

***Abstract*—It's challenging to automatically produce brief descriptions of an image's meaning because it can have diverse connotations in different languages. However, due to the vast amount of information packed into a single image, it is challenging to parse out the necessary context to use it to build sentences. It's a great way for the visually impaired to get around independently. This type of system can be built using the emerging programming technique of deep learning.**

This paper presents the development of an Image Caption Bot designed to aid individuals with visual impairments. We achieve enhanced accuracy in caption generation by modeling on the MSCOCO dataset using a Transformer encoder and Inception v3 for image processing. Image captioning, which entails generating textual descriptions for images, is the primary focus of our research. We achieve enhanced accuracy in caption generation by utilizing a Transformer encoder during training. The MSCOCO dataset serves as a valuable

The results of the model are translated into speech for the benefit of the visually handicapped.

Keywords*—*CNN, Google Text To Speech, MS-COCO, Inspection v3.

1. Introduction

Globally, approximately 1.3 billion individuals grapple with vision impairments, either near or distant, which can hinder their ability to navigate through rooms or along various travel routes. The primary objective of this project is to aid visually impaired individuals who encounter obstacles in perceiving objects. The implementation of this project has significantly enhanced the quality of life for the visually impaired, enabling them to independently navigate their surroundings without the need for assistance. They are now capable of managing their affairs and traversing streets unaided. The realm of computer vision, particularly in the context of neural networks, has seen significant advancements. The aim is to utilize the output of a controller to guide visually impaired individuals. The project's objective is to develop a system that will assist visually impaired individuals in their day-to-day activities. These technologies empower visually impaired individuals to work more efficiently, eliminating the need for dependence on others to perform tasks.

2. Literature Review

[1] This paper delves into the operational components utilized in object detection, specifically focusing on Object Detection, Traffic Sign Detection, and Voice Assistant functionalities. The Smart Blind Navigating system addresses these functionalities by pTensorFlowetailed and contextually rich information about the user's surroundings, facilitating navigation, and enhancing overall system accuracy while ensuring user safety. Two algorithms are discussed: 1) The TensorFlow algorithm, is used for object detection in videos and photos through an API. 2) The R-CNN algorithm, encompassing object detectors such as R-CNN, Fast R-CNN, and Faster R-CNN. The system utilizes Deep Learning techniques to identify objects, functioning effectively under diverse environmental conditions.

[2] This paper explores existing methodologies in object detection, including the YOLO model, Point detection and tracking, and Segmentation detection and tracking. By analyzing the needs of blind and

visually impaired individuals, the study aims to develop a blind visualization system inspired by Convolutional Neural Networks (CNNs). This system aims to assist blind individuals in navigating their surroundings more effectively by offering a portable, real-time solution. It leverages mobile cameras, high-definition video links, and robust server infrastructure to generate 3D sound representations.

[3] The rapid expansion of the automotive industry and evolving market demands necessitate the development of new algorithms and solutions to enhance performance and features. This study addresses the need for alternatives to LIDARs for distance estimation, advocating for the use of cameras. The proposed approach relies on deep neural networks, particularly the YOLO model, and stereoscopy concepts. It employs two offset cameras to capture images, which are then processed using stereoscopic algorithms to estimate distances to detected objects.

[4] This paper outlines a strategy for improving YOLO's ability to predict object distances accurately using data solely from a monocular camera. The study introduces class-aware and class-agnostic methods for handling distance estimation, demonstrating that the class-agnostic approach yields smaller prediction vectors and superior outcomes. The integration of distance measurement and object detection tasks enhances the accuracy of bounding box predictions. Recent iterations of YOLO, including versions four and five, emphasize efficiency improvements while retaining the core architecture. The newest iteration, YOLOX, builds upon YOLOv3, introducing an anchor-free algorithm that facilitates independent usage of detection heads. Experimental results demonstrate the system's effectiveness, achieving an average relative error of 11% across eight classes and a distance range of 0 to 150 meters using the KITTI dataset.

[5] Movement Estimation Using Mediapipe BlazePose is a technique described in this paper, which involves analyzing body movements captured by video cameras and overlaying skeleton joints with labels onto the subject's body. The study emphasizes the real-world applications of this research, particularly in physically demanding workplaces and the sports industry. Deep learning algorithms are utilized to identify joints within a person's body, with Mediapipe Blazepose being the chosen algorithm for detecting and analyzing movements intended to cause bodily harm under stress. A comparison with IMU-based motion capture indicates less than 1a 0% accuracy difference, with deep learning methods leveraging actual sensor data for analysis, as opposed to 2D image analysis by IMUs. The ultimate goal of this project is to develop a functional system capable of accurately identifying skeleton joints and calculating movement velocity and joint angles, aiding in the assessment of potential injury risk associated with certain body movements.

[6] This paper introduces a new system based on posture detection, aiming to address the increasing prevalence of remote work during the COVID-19 pandemic. Improper desk heights, inadequate ergonomic equipment, and prolonged laptop usage have been identified as contributors to poor posture, resulting in various musculoskeletal issues such as back pain and neck strain. By utilizing posture detection technology, the proposed system seeks to mitigate these health concerns and improve user well-being, particularly for individuals working extended hours in remote settings.

[7] This paper delves into the critical endeavor of improving text-to-speech technology to imbue it with greater expressiveness and naturalness. By shedding light on the deficiencies inherent in current text-to-speech systems, which frequently yield robotic-sounding voices devoid of emotional nuance, the study underscores the pressing need for advancements in this domain. Through meticulous analysis and innovative approaches, the research aims to bridge the gap between artificial speech synthesis and human-like vocal expression, enhancing the overall quality and user experience of text-to-speech systems. This endeavor holds promise for various applications, including virtual assistants, accessibility tools, and entertainment media, where lifelike vocalizations are paramount for effective communication and engagement.

[8] This study describes the development of a phoneme-based text-to-speech system for English, utilizing a cascade-parallel formant synthesizer. The system comprises language processing, auditory processing, and synthesis components, enabling the conversion of text input into phoneme strings with added prosodic information.

3. Algorithms

- **Convolutional Neural Networks (CNNs) and Inception v3 for Image Analysis:**

Convolutional Neural Networks (CNNs) have revolutionized the field of computer vision by enabling automated feature extraction from images. However, the design of CNN architectures often involves trade-offs between computational efficiency and model performance. Inception v3, a variant of the Inception architecture, addresses these challenges by introducing innovative design choices that enhance both efficiency and accuracy.

Incorporating both CNNs and Inception v3 into image analysis workflows presents a powerful approach. Initially, CNNs serve as the backbone for feature extraction, leveraging their ability to capture hierarchical representations of image content. This initial feature extraction step provides a foundation for subsequent analysis.

Inception v3 then complements CNNs by introducing advanced architectural elements such as the Inception module. This module utilizes parallel convolutional layers with varying filter sizes, enabling the network to capture multi-scale features effectively. Additionally, Inception v3 incorporates factorization techniques to reduce computational complexity, making it suitable for deployment in resource-constrained environments.

- **Google Text-to-Speech (GTTS):**

Text may be converted into a natural-sounding voice with Google Text-to-voice (TTS), a crucial part of digital communication and accessibility. TTS is powered by sophisticated neural network models and has a strong synthesis engine. Google TTS enables inclusive design in a variety of applications, such as assistive technology, navigational systems, and language learning tools, by supporting numerous languages and diverse voices.

Developers are enabled to innovate and improve user experiences through its integration via APIs. Future developments in text-to-speech technology will be facilitated by addressing issues like voice quality and pronunciation accuracy, even though Google TTS greatly increases accessibility and usefulness.

4. Terminologies

- **Image Captioning:**

Image Captioning is a procedure that amalgamates both computer vision and natural language processing to formulate a textual depiction of an image. This process typically employs an encoder-decoder framework: The encoder scrutinizes the input image and encodes it into a set of features, thereby creating an intermediate representation of the image.The decoder, which is often a language model, subsequently generates a textual description based on these features.

This technology can be particularly beneficial for individuals with visual impairments, as it enables them to comprehend the content of images. Moreover, it finds usage in various applications such as content moderation, image indexing, and more.

- **Transformer Encoder/Decoder:**

The Transformer model, utilized for tasks such as machine translation, incorporates an encoder-decoder structure. The encoder processes the input sequence into a continuous representation or embedding. It is composed of multiple identical layers, each containing two sublayers: a multi-head self-attention mechanism and a fully connected feed-forward network.The decoder employs these embeddings from the encoder to generate the output sequence. It also consists of multiple identical layers. Each decoder block receives the

features from the encoder.

Fundamentally, the encoder maps an input sequence to a sequence of continuous representations. The decoder then uses these, along with its previous outputs, to generate an output sequence.This architecture enables the model to effectively handle long sequences and capture complex patterns within the data. It serves as the foundation for many state-of-the-art models in natural language processing, such as BERT and GPT-3.

- **Inception v3:**

Inception v3 is a convolutional neural network (CNN) architecture developed by Google Research, belonging to the Inception family of models. It is specifically designed for image classification tasks and has achieved state-of-the-art performance on various benchmarks. Inception v3 is trained on large-scale datasets such as the ImageNet dataset, which contains millions of labeled images across thousands of categories.

The architecture of Inception v3 is characterized by its innovative design, featuring inception modules that incorporate parallel convolutional layers with different filter sizes. This design allows the network to capture multi-scale features effectively, enhancing its ability to understand complex visual information. Inception v3 also incorporates factorization techniques to reduce computational complexity while maintaining accuracy. Overall, Inception v3 has become a widely used model in the field of computer vision, providing a powerful tool for tasks such as image recognition, object detection, and image captioning.

- **MS-COCO Dataset:**

The MSCOCO Dataset, important in computer vision and understanding language, includes images with many captions. Annotations match words with pictures, helping teach models. Captions give details, like what objects are and what's happening in scenes. MSCOCO has lots of different things to look at, like different objects and places. Image Description Generation means making captions without help, using MSCOCO's notes. Object Recognition is finding objects in pictures, helped by MSCOCO's notes about where things are. Semantic Understanding is when models know what things in pictures mean. Contextual Comprehension is understanding how things in pictures relate. This helps explain important ideas in using MSCOCO for making captions for pictures.

5. Implemented System

- **Object detection:**

Following an in-depth analysis of two prominent algorithms and synthesizing common insights gleaned from surveys, our focus shifts towards the implementation of a system aimed at enhancing navigation for the visually impaired. This system is poised to provide directional guidance, thereby improving the travel experience for individuals with visual impairments. Leveraging object detection technology, we envision empowering users by training the camera to identify a wider array of objects. This enhancement is anticipated to enable safer navigation across diverse neighborhoods.

- **Text To Speech Conversion:**

Our investigation into text-to-speech conversion techniques has led us to adopt Google Text-to-Speech (GTTS). This versatile tool provides strong capabilities for converting text into speech across various languages. While our efforts include support for multiple languages, our primary focus remains on facilitating smooth conversion in English. By using the text-to-speech synthesis algorithm and GTTS, we aim to create a complete solution that makes sure the spoken words sound natural in various languages.

Seamless Integration: GTTS seamlessly integrates into our existing framework, ensuring efficient processing and delivery of synthesized speech output.

- **Implemented architecture / System diagram:**

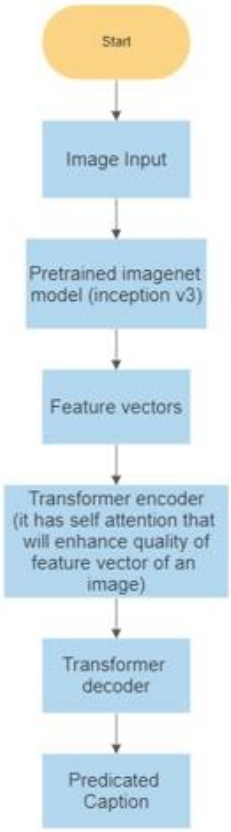

Figure 5.1: Flow diagram of Caption Bot

- **Plan of activation**

1. **Image Input:** Capture an image as the input to the system.
2. **Object Detection and Classification:** Calculate the distance of objects in the image and identify their classes using an object detection and classification algorithm.
3. **Description Generation:** Generate a textual description of the scene based on the detected objects and their positions. This description will be in the form of a string.
4. **Text-to-Speech Conversion:** Pass the generated string description to a text-to-speech converter to convert it into audio.
5. **Speech Synthesis:** Simultaneously perform speech synthesis to convert the text into human-like speech.
6. **Output Generation:** Produce the final output in the form of synthesized speech, which provides an audible description of the scene captured in the input image.

6. Conclusion

In our exploration and implementation of the "Image Caption Bot For Assistive Vision" system, we have meticulously designed a framework that integrates cutting-edge technologies to cater to the needs of visually impaired individuals. By leveraging Object Detection through Convolutional Neural Networks (CNN) with the powerful Inception V3 architecture. Additionally, employing a transformer encoder-decoder model with multi-attention mechanisms enhances the system's capability to generate descriptive and contextually relevant captions.
Furthermore, the integration of Text-to-Speech Conversion using Google Text-to-Speech (GTTS) enables seamless auditory presentation of these captions. Through an examination of diverse datasets like MS COCO and ImageAI, we have enriched our understanding and refined our approach. By combining these modules, we anticipate the creation of an assistive system that not only bridges accessibility gaps but also empowers visually impaired individuals to engage with their surroundings more independently and effectively, thereby significantly improving their quality of life.

7. References

[1] Pooja Maid, Omkar Thorat, Sarita Deshpande , “Object Detection for Blind User’s ” International Research Journal of Engineering and Technology (IRJET) ,Volume: 07, June | 2020

[2] N.Saranya , M.Nandinipriya , U.Priya , “Real Time Object Detection for Blind People ” ,Bannari Amman Institute of Technology, Sathyamangalam, Erode.(2018).

[3] Bojan Strbac, Marko Gostovic, Zeljko Lukac “ YOLO Multi-Camera Object Detection and Distance Estimation ” , 2020 Zooming Innovation in Consumer Technologies Conference (ZINC),May 2020.

[4] Maxwell Abedi, Dan.O.M.Bonsu, Isaac K. Badu, Richmond Afoakwah, Pooja Ahuja

"Spectroscopic (analytical) approach to gunshot residue analysis for shooting distance estimation: a systematic review",September 2020.

[5] Pauzi, A.S., Nazri, F.B., Sani, S., Bataineh, A.M., Hisyam, M.N., Jaafar, M.H., Wahab, M.N., & Mohamed, A.S. (2021). Movement Estimation Using Mediapipe BlazePose. IVIC.

[6] Ananya Ashok Naik, Krishna, Neha Kishor Mudhol, Qamrath Akthar Sheikh, Mr. Ramesh Nayak, "Sitting Posture Monitoring System using Image Classification" (2022).

[7] John F. Pitrelli, Member, IEEE, Raimo Bakis, Ellen M. Eide, Raul Fernandez, Wael Hamza, and Michael A. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English" , IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 4, JULY 2006 .

[8] " A Multilingual Text-to-Speech System" Author : Hector Javkin.

[9] Ashwani Kumar , Sonam Srivastava , " Object Detection System Based on Convolution Neural Networks Using Single Shot Multi-Box Detector " , Third International Conference on Computing and Network Communications (CoCoNet'19).

[10] May year 2010. Chen X, Ho CT, Lim ET, Kyaw TZ ,Cellular Phone based online ECG Processing for Ambulatory and Continuous Detection. Computers in Cardiology, volume 34, page no 653-656, year 2008.

[11] Hanen Jabnoun, Faouzi Benzarti,and Hamid Amiri, "Object recognition for blind people based on features extraction " IEEE IPAS14: INTERNATIONAL IMAGE PROCESSING APPLICATIONS AND SYSTEMS CONFERENCE 2014.

[12] Yun J, Lee SS. Human movement detection and identification using pyroelectric infrared sensors. Sensors (Basel). 2014 May 5;14(5):8057-81. doi: 10.3390/s140508057. PMID: 24803195; PMCID: PMC4063065.

[13] Anand Upadhyay, Ketan Chaudhari, Pradip Bhere, Jerin Thomas, "Body Posture Detection Using Computer Vision," SSRG International Journal of VLSI & Signal Processing, vol. 7, no. 1, pp. 6-10, 2020.

[14] Nguyen, Phat & Thi, Ngoc & Ngoc, Thien. (2021). Proposing Posture Recognition System Combining MobilenetV2 and LSTM for Medical Surveillance. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3138778.

[15] Zhang, K.; Xie, J.; Snavely, N.; Chen, Q. Depth sensing beyond lidar range. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,
Seattle, WA, USA, June 2020.

[16] Natanael, G.; Zet, C.; Foşalău, C. Estimating the distance to an object based on image processing. In Proceedings of the 2018 International Conference and Exposition on Electrical Furthermore, Power Engineering (EPE), Iasi, Romania, 18–19 October 2018.

[17] Huang, L.; Zhe, T.; Wu, J.; Wu, Q.; Pei, C.; Chen, D. Robust inter-vehicle distance estimation method based on monocular vision. IEEE Access 2019.

[18] Luo, X.; Huang, J.B.; Szeliski, R.; Matzen, K.; Kopf, J. Consistent video depth estimation. ACM Trans. Graph. (TOG) 2020.

[19] Itunuoluwa Isewon , Jelili Oyelade , Olufunke Oladipupo "Design and Implementation of Text To Speech Conversion for Visually Impaired People" , International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, April 2014 .