Image Captioning tool to Enhance The Virtual Visibility of Partial Blind People

*

Piyush Chaudhary
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
chaudhary.piyush0801@gmail.com

Ritvik Rahotgi
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
rohatgiritvik@gmail.com

Prachi Verma
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
2000prachiverma12513@gmail.com

Naveen Chauhan
Assistant professor
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
naveen.chauhan@kiet.edu

***Abstract*—Every day, more than 10 million people navigate their surroundings without the sense of sight. IMate offers a solution to this challenge using Image Captioning technology, which utilizes advanced deep learning techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) to identify and describe images in English. Image captioning is a Sequence to Sequence problem that combines natural language processing and computer vision concepts. To accomplish this, both language processing and image feature extraction are necessary, which can be achieved using NLP and CNN, respectively. With a vast dataset and the power of deep learning, IMate provides an innovative and practical solution for the visually impaired. This service is unique and not just another dime-a-dozen.**

***Index Terms*—Dime-a-dozen,services, MERN, DOM**

I. Introduction

The visually impaired community faces various obstacles in their daily lives, particularly when it comes to navigating and understanding their surroundings. The lack of sight can significantly hinder their independence and autonomy. To address this issue, researchers and technologists have been developing innovative solutions to help the visually impaired move around more easily. One such solution is the IMate, a device that employs advanced deep learning techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) to recognize and describe images.

IMate is an image captioning technology that provides the visually impaired with a description of their environment. Image captioning uses natural language processing and computer vision concepts to convert images into English descriptions. To achieve this, both language processing and image feature extraction are necessary, which can be done using NLP and CNN, respectively. IMate can process vast amounts of visual data to provide accurate and comprehensive descriptions of the environment to the visually impaired.

The capabilities of IMate are made possible by state-of-the-art deep learning techniques, including Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM). CNN is particularly suitable for image recognition tasks, as it can extract relevant features from images and classify them accurately. LSTM is a recurrent neural network that is ideal for sequence prediction tasks, making it perfect for language processing. By combining these two techniques, IMate can accurately recognize and caption images.

The primary objective of IMate is to provide the visually impaired with a practical and effective solution for navigating their environment. With the power of deep learning and a vast dataset, IMate can identify and describe a wide range of objects, environments, and situations. By providing accurate and comprehensive descriptions, IMate can help the visually impaired move around their surroundings with greater confidence and independence.

Overall, IMate is an innovative solution that utilizes cutting-edge deep learning techniques to address the challenges faced by the visually impaired. With its ability to identify and caption images, IMate has the potential to significantly improve the quality of life for the visually impaired, providing them with greater autonomy and independence.

II. RELATED WORKS

IMate is a technology designed to assist visually impaired individuals by providing descriptions of their surroundings using advanced image captioning techniques. The system uses deep learning methods such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) to analyze images taken by the user and generate corresponding English descriptions.

The process starts with the user taking a picture of their environment with the device's camera. IMate then uses its deep learning algorithms to extract relevant features from the image, including objects, shapes, and colors. These features are inputted into the LSTM model, which generates an accurate and comprehensive description of the image.

The description is then converted into speech and played back to the user through a text-to-speech system. This allows visually impaired individuals to understand their surroundings and navigate safely without relying on sight. IMate is not limited to indoor environments and can also be used outdoors to identify and describe buildings, streets, and landmarks.

In summary, IMate provides a practical solution for visually impaired individuals to navigate their environment independently. By using advanced deep learning techniques, the system accurately identifies and describes images, providing users with the information they need to navigate their surroundings with confidence.

III. BACKGROUND TOOLS

A. Frontend - HTML, CSS, JavaScript.

HTML, also known as Hypertext Markup Language, is the industry standard markup language for creating web pages. It enables you to define the structure of a web page by incorporating headings, paragraphs, images, and links. In the context of IMate, HTML can be utilized to develop the user interface by designing the layout of various components and placing interactive elements such as buttons and input fields.

Cascading Style Sheets, or CSS, is a stylesheet language utilized to specify the presentation of a document composed in HTML. It enables you to define the visual aspects of the UI, including colors, fonts, and styling of different components. CSS aids in creating an appealing and user-friendly UI for IMate that is easy to navigate.

JavaScript is a programming language employed to create dynamic and interactive web pages. It can be utilized to add functionality to the UI, such as capturing images from the camera, processing them with the IMate algorithms, and displaying the generated descriptions to the user. JavaScript is instrumental in creating a responsive and user-friendly UI that provides real-time feedback and a captivating user experience.

By combining HTML, CSS, and JavaScript, you can create an aesthetically pleasing and functional UI for IMate. The UI can be designed to be inclusive for visually impaired individuals with features such as high contrast, large fonts, and screen readers. Additionally, the UI can be optimized for various devices such as smartphones, tablets, and desktops, ensuring a consistent user experience across all platforms.

B. Backend – Flask

In the case of IMate, Flask can be used to build the backend API that handles image processing and caption generation. Flask can handle HTTP requests sent from the frontend and provide appropriate responses to them. For example, when the user captures an image using the frontend UI, the frontend can send an HTTP POST request to the Flask backend containing the image data.

The Flask backend can then process the image using the advanced deep learning techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) to identify and caption the image. The generated caption can then be sent back to the frontend as an HTTP response, which can be displayed to the user.

Flask provides many benefits for building the backend of IMate, such as:

Ease of use: Flask is easy to learn and use, with a minimalistic approach that focuses on simplicity and flexibility.

Modularity: Flask allows developers to create small, modular components that can be combined to create larger applications.

Flexibility: Flask can be used with a wide range of databases, libraries, and other tools, making it a versatile framework for building web applications.

Scalability: Flask can handle large volumes of traffic and can be scaled easily to meet the demands of a growing user base.

Overall, Flask provides a powerful and flexible backend framework that can be used to build the API for IMate, allowing the frontend UI to communicate with the image processing and caption generation algorithms.

C. Database – Flicker30k dataset.

The Flicker30k dataset is a collection of 31,000 images and 158,000 captions in English, obtained from the Flickr website. It encompasses a wide range of topics, including people, animals, landscapes, and objects, and features multiple captions for each image, each of which describes a different aspect of the picture.

As a widely used dataset in image captioning research, the Flicker30k dataset has played a pivotal role in developing state-of-the-art deep learning models for this task. Its size and diversity enable the training and testing of models that can handle various images and scenarios.

To incorporate the Flicker30k dataset into IMate, the images and captions can be preprocessed and input into deep learning models using methods like Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM). These models can then be trained on the dataset to learn how to generate captions for new images fed into the system.

In summary, the Flicker30k dataset offers a valuable resource for training deep learning models that power IMate's image captioning capabilities. With its vast and diverse collection of images and captions, the dataset can help develop

D. Machine Learning – Image Processing , NLP, CNN, Transfer Learning, LSTM.

Machine Learning techniques used in IMate:

Image Processing: Image processing is a technique used to extract meaningful information and features from input images. In the field of image processing, Convolutional Neural Networks (CNNs) are a widely used deep learning model for image recognition and classification. To perform image processing in IMate, a pre-trained CNN model such as VGG or ResNet is used to extract features from images. These models are trained on large-scale image recognition datasets, which allows them to recognize and classify images accurately.

NLP:IMate employs natural language processing (NLP) techniques to produce natural language descriptions of input images. Specifically, the system uses a variant of NLP known as sequence-to-sequence (Seq2Seq) learning. This approach involves two recurrent neural networks (RNNs): an encoder network that processes the input image and a decoder network that generates the output caption. The Seq2Seq models are trained on a vast collection of image-caption pairs, such as those found in the Flicker30k dataset.

Transfer Learning: Transfer learning is a method that involves using a pre-existing deep learning model to start a new task that is related. In the context of image processing in IMate, transfer learning is utilized to fine-tune a pre-trained CNN model so that it is better suited for the task of generating natural language descriptions of images. This involves modifying the pre-existing CNN model and training it on a portion of the Flicker30k dataset to acquire image features that are pertinent to the image captioning objective.

LSTM: LSTM (Long Short-Term Memory) networks belong to the family of RNNs (Recurrent Neural Networks) and are capable of processing longer sequences of data, such as sentences in an image caption. Within the context of IMate, LSTM networks are utilized in the Seq2Seq model for generating natural language descriptions of images. These networks are especially advantageous for this task since they can be trained to retain crucial information from the beginning of the input sequence and use it to produce more precise and logical captions.

IMate utilizes various Machine Learning methods such as Image Processing, NLP, Transfer Learning, and LSTM to furnish individuals with visual impairments with natural language explanations of their surroundings. These methods extract features from images, create natural language descriptions, and fine-tune pre-existing models to suit the image captioning task.

E. Additional Hardware/ Paid Software Required – Raspberry pi, Basic Camera

Hardware required for IMate:

Raspberry Pi: The Raspberry Pi is a single-board computer that is compact, cost-effective, and well-suited for running IMate's software. Due to its processing power and memory capabilities, it's often used for deep learning applications. Additionally, it's user-friendly and affordable. The Raspberry Pi is typically connected to a display device, like a monitor or tablet, to show the image captions that IMate generates.

Basic Camera: To use IMate for image captioning, you will need a camera to capture images of the environment. A basic camera can be connected to a Raspberry Pi to capture real-time images. However, it's important to keep in mind that the quality of the camera can have an impact on the accuracy of the image captioning. Therefore, it is recommended to use a high-resolution camera with good image quality to achieve the best results.

The IMate system consists of two main hardware components: the Raspberry Pi and a basic camera. The Raspberry Pi is responsible for providing the necessary processing power and memory to run the software. On the other hand, the camera captures images that are processed by the deep learning models. By working together, these hardware components enable the IMate software to provide natural language descriptions of the environment, which can assist visually impaired individuals in navigating and comprehending the world around them.

IV. IMATE WORKS USING THE VARIOUS TECHNOLOGIES INVOLVED:

A. Image processing

IMate utilizes image processing methods to detect and examine the images taken by the camera of the device. The initial image is subjected to preprocessing procedures to eliminate any undesired components, such as noise, that could impede the process of generating image captions.

B. Convolutional Neural Networks (CNN)

IMate utilizes a pre-trained Convolutional Neural Network (CNN) model, such as VGG 16, to extract significant features from the input image. This CNN model is trained on an extensive dataset of images to comprehend how to recognize various objects, individuals, and other visual components in the image. By obtaining these features, IMate is capable of producing a thorough and precise depiction of the image.

C. Natural Language Processing (NLP)

IMate employs NLP techniques to generate a natural language description of an image after extracting its features. This process includes text pre-processing, tokenization, and word embeddings to transform the image features into a word sequence that depicts the image.

D. Long Short-Term Memory Networks (LSTM)

IMate employs LSTM neural networks to produce natural language descriptions of images. LSTM is a kind of recurrent neural network that can handle data sequences. In the context of IMate, the LSTM utilizes the sequence of image features created by the CNN, as well as a beginning sequence token, to produce a sequence of words describing the image.

E. Transfer Learning

IMate utilizes transfer learning methodologies to enhance the precision and efficiency of its image captioning process. In transfer learning, a pre-existing model, such as VGG 16, is employed and fine-tuned for a specific task, such as image captioning. By fine-tuning the pre-trained model, IMate is able to achieve superior performance even with a smaller training dataset.

F. Flask

The backend of IMate employs Flask, a Python web framework that facilitates handling user requests and communication with the image processing, CNN, and LSTM models. Flask is known for its lightweight and versatile nature, which enables seamless integration with various technologies.

G. Raspberry Pi

IMate has chosen the Raspberry Pi as the hardware platform for their device due to its compact size, cost-effectiveness, and energy efficiency. This single-board computer is particularly suitable for executing the image processing and machine learning algorithms that IMate employs.

H. HTML, CSS, and JavaScript

IMate uses HTML, CSS, and JavaScript for the frontend user interface. The frontend provides an easy-to-use interface for users to interact with the device and receive real-time descriptions of their environment.

Overall, IMate is a solution designed to help visually impaired individuals understand their surroundings. It combines advanced techniques such as image processing, machine learning, and natural language processing. Through the use of deep learning techniques like CNN and LSTM, IMate is capable of providing real-time, accurate, and detailed descriptions of the user's environment. The system also leverages pre-trained models and transfer learning to enhance its capabilities.

V. Future Works

A. Dataset Expansion

To enhance the accuracy and versatility of IMate, future research could consider incorporating a more diverse range of images and environments into the dataset beyond the current Flickr 8K dataset, which is limited in size.

B. Improved Hardware

While the Raspberry Pi can serve as a viable hardware platform for IMate, there is potential for further research to investigate the benefits of utilizing more advanced and powerful hardware, such as GPUs, in order to enhance the speed and overall performance of the system.

C. Multi-Language Support

IMate currently provides descriptions only in English. However, it could be worthwhile for future research to explore the possibility of introducing multi-language support. This would enable users to receive descriptions in the language they prefer.

D. Real-Time Object Recognition

At present, IMate is capable of describing a user's surroundings, but it is limited to providing general information and does not have the ability to identify individual objects within the environment. To enhance the accuracy and level of detail in the descriptions produced by the system, it may be beneficial to investigate the integration of real-time object recognition technologies in future research.

Overall,IMate shows great promise as a solution for assisting visually impaired individuals in navigating their surroundings through the use of advanced image processing, machine learning, and natural language processing techniques. As further research and development are conducted, IMate could become an even more effective and popular solution for enhancing the quality of life for visually impaired individuals.

VI. Conclusion

IMate is a solution designed to help individuals who are visually impaired. It employs advanced technologies such as machine learning, natural language processing, and image processing to enable users to understand their surroundings better. IMate uses deep learning methods like CNN and LSTM to generate real-time, accurate descriptions of the user's environment. The solution leverages pre-trained models and transfer learning techniques to achieve better results with less training data, making it cost-effective. With a Raspberry Pi and a user-friendly frontend interface, IMate provides a practical, accessible, and affordable solution for visually impaired individuals to navigate their surroundings with more independence and confidence.

Acknowledgment

The authors would like to express their appreciation and thanks to the Department of Computer Science and Engineering, **KIET Group of Institutions**,Ghaziabad,India for their immense support.

References

[1] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 1097-1105. (2012)

[2] Girshick, Ross, et al. "Region-based Convolutional Networks for Accurate Object Detection and Segmentation." IEEE Transactions on Pattern Analysis Machine Intelligence 38.1:142-158. (2015)

[3] Devlin, Jacob, et al. "Language Models for Image Captioning: The Quirks and What Works." Computer Science (2015)

[4] Fang, H., et al. "From captions to visual concepts and back." Computer Vision and Pattern Recognition IEEE, 1473-1482. (2015)

[5] Cho, Kyunghyun, et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." Computer Science (2014)

[6] Hochreiter, Sepp, and J. Schmidhuber. "Long Short-TermMemory."Neural Computation 9.8: 1735-1780. (1997)

[7] Karpathy, Andrej, and F. F. Li. "Deep visual-semantic alignments for generating image descriptions." Computer Vision and Pattern Recognition IEEE, 3128-3137. (2015)

[8] Sermanet, Pierre, et al. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks." Eprint Arxiv (2013)

[9] Sundermeyer, M., et al. "Comparison of feedforward and recurrent neural network language models." IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 8430-8434. (2013)

[10] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." Computer Science (2014)
[11] Szegedy, Christian, et al. "Going deeper with convolutions." IEEE Conference on Computer Vision and Pattern Recognition IEEE, 1-9. (2015)
[12] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 770-778. (2016)
[13] Mao, Junhua, et al. "Explain Images with Multimodal Recurrent Neural Networks." Computer Science (2014)
[14] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 3156-3164. (2015)
[15] Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." Computer Science ,2048-2057. (2015)
[16] Papineni, K. "BLEU: a method for automatic evaluation of MT." (2001)