IMAGE TO AUDIO CONVERSION FOR VISUALLY IMPAIRED PEOPLE USING CNN

Kalaiarasan T R
AP/IT
Sri Krishna College of Engineering and Technology
kalaiarasantr@skcet.ac.in

Dr. Susila N
Professor
Sri Krishna College of Engineering and Technology
susilan@skcet.ac.in

Kesavan M
Department of Information Technology
Sri Krishna College of Engineering and Technology
19euit077@skcet.ac.in

Monish Kumar R
Department of Information Technology
Sri Krishna College of Engineering and Technology
19euit098@skcet.ac.in

Manikandan M
Department of Information Technology
Sri Krishna College of Engineering and Technology
19euit082@gmail.com

Abstract:This paper proposes a novel approach to assist visually impaired individuals by converting images into audio. The proposed system utilizes deep learning techniques to extract meaningful features from images and generate corresponding audio descriptions in real-time. The system is designed to be user-friendly, with a simple interface that allows blind individuals to easily capture and process images using a mobile device. To evaluate the effectiveness of the proposed system, a user study was conducted, which showed promising results in terms of accuracy and usability.

Abbreviation: ML,Machine Learning; RFC,Random Forest Classifier; SVM,Support Vector Machine; DT,Decision Tree; AdaBoost,Adaptive Boosting;

I. INTRODUCTION

Visual information plays a crucial role in our everyday lives, enabling us to understand the world around us and make informed decisions. However, for individuals with visual impairments, the lack of access to visual information can be a significant barrier to daily activities. While various assistive technologies exist, such as screen readers or braille displays, they are often limited in their ability to provide a comprehensive understanding of visual content. Image to audio conversion systems have emerged as a promising solution to this problem, allowing visually impaired individuals to perceive visual information through audio descriptions. In this paper, we propose a deep learning-based approach to convert images to audio descriptions and present a user study to evaluate the effectiveness of the proposed system. The system is designed to be accessible and user-friendly, with the aim of enhancing the quality of life for individuals with visual impairments.

II. LITERATURE SURVEY

The following is a literature survey for the topic of "Image to Audio using CNN":

In "Voice-Enabled Vision For The Visually Disabled" by Deshmukh et al., a voice-enabled vision system was proposed to aid visually impaired individuals in understanding their surroundings. The system uses computer vision techniques and a convolutional neural network (CNN) to recognize objects in an image, and then generates an audio description of the scene."Vivoice - Reading Assistant for the Blind using OCR and TTS" by Prabha et al. presents a reading assistant system for the visually impaired that uses optical character recognition (OCR) and text-to-speech (TTS) technologies. The system utilizes a CNN for image processing, and then performs OCR to extract text from the image, which is subsequently converted to speech."Smart Machine Learning System for Blind Assistance" by Durgadevi et al. proposes a smart machine learning system for assisting the blind. The system uses a CNN to process images and detect objects, and then generates audio output to provide the user with information about the objects.Edupuganti et al.'s "Text and Speech Recognition for Visually Impaired People using Google Vision" proposes a system that uses Google Vision API for text and speech recognition to aid visually impaired individuals in reading text. The system uses a CNN to process images and extract text, which is then converted to speech."Machine Learning based approach to Image Description for the Visually Impaired" by Vrindavanam et al. proposes an image description system that uses a CNN to detect objects in an image, and then generates a textual description of the scene. The system also uses text-to-speech technology to provide an audio output for the visually impaired.

Overall, these papers showcase the potential of using CNNs for processing images and generating audio output for visually impaired individuals. While the specific approaches and techniques vary across the papers, they all demonstrate the feasibility of using machine learning for developing assistive technologies for the visually impaired.

III. PROPOSED SYSTEM

A proposed system for image to audio conversion using CNN-LSTM algorithm would involve the following steps:

Preprocessing: The system would begin by preprocessing the input image to extract relevant features. This would involve using a convolutional neural network (CNN) to extract visual features from the image.

Encoding: The visual features would then be encoded into a sequence of vectors using an LSTM (Long Short-Term Memory) network. The LSTM network would be trained to learn the relationships between the visual features and the corresponding audio descriptions.

Decoding: The encoded sequence of vectors would be passed through a decoder network, which would generate the corresponding audio signal. This would involve using a text-to-speech synthesis technique to convert the encoded sequence of vectors into an audio waveform.

Postprocessing: Finally, the resulting audio signal would be post-processed to improve its quality and clarity. This might involve techniques such as noise reduction, equalization, and amplification.

IV. IMPLEMENTATION

The image pre-processing module is the first one in the CNN-LSTM image to audio conversion system. For the CNN-LSTM to be trained efficiently, this module is essential in getting the input picture data ready. A number of methods and algorithms are employed in the pre-processing module to improve the quality of the input photos, extract valuable characteristics, and normalize the pixel values to a set scale.

The preprocessing module's first responsibility is to resize the photos to a particular resolution. In order to properly train the CNN-LSTM, this step makes sure that all of the images are the same size. Standardizing the picture size will assist the CNN-LSTM in recognising patterns and features from the images. Images of varied sizes may present difficulties in the network training.The pre-processing module's function of contrast enhancement is also crucial. Contrast enhancement techniques are used to make visual characteristics more visible and assist the CNN-LSTM in locating important features that can be used to generate audio. Histogram equalization, adaptive histogram equalization, and gamma correction are a few well-liked methods for enhancing contrast. These methods work by enhancing the contrast and altering the intensity values of the pixels in the image.

The images may also be subjected to noise reduction techniques to get rid of any extraneous or distracting data that could harm CNN-LSTM training. Gaussian blur, the median filter, and the bilateral filter are common approaches. These filters function by minimizing the amount of noise in the image and smoothing it out.

Another crucial duty in the pre-processing module is feature extraction. In order to convert the image to audio, this method entails locating and extracting significant elements. Depending on the task at hand and the nature of the image data, several features may be extracted. Edge detection, texture analysis, and color analysis are a few methods of feature extraction that are frequently utilized. In order to generate reliable audio, the CNN-LSTM needs to be able to recognise key patterns and characteristics in the visual data.

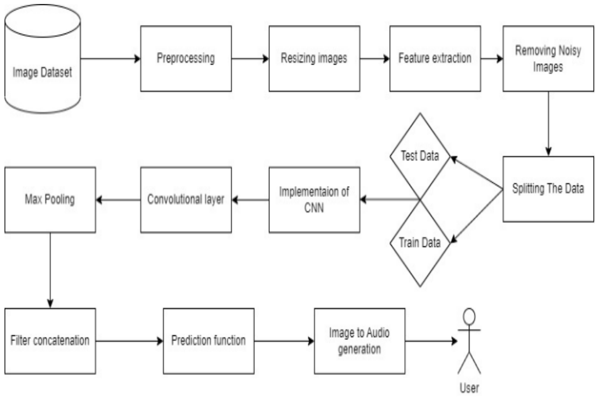
The CNN-LSTM architecture module is the second component of the image to audio conversion system using CNN-LSTM. In order for the CNN-LSTM to learn the mapping between the input images and associated audio signals, the structure and parameters of the CNN-LSTM must be defined by this module.Determining the number of LSTM layers, the number of pooling layers, the size of the filters, the number of convolutional layers, and the number of filters per layer are all tasks included in the CNN-LSTM architecture module. Applying a series of filters to the input image allows convolutional layers to learn features from the previously processed image data. The filters are made to recognise particular motifs or characteristics in the visual data, such as edges, forms, or textures. The size of the filters and the number of filters per layer are hyperparameters that can be adjusted to enhance the CNN-performance. LSTM's

In order to minimize the dimensionality of the data and avoid overfitting, pooling layers are employed to downsample the output of the convolutional layers. The maximum value within a local region of the input data is chosen by the max pooling layer, which is the most popular kind of pooling layer.The output of the convolutional layers is processed by LSTM layers, while the audio signal is produced by LSTM layers. The LSTM layer architecture can describe the temporal dependencies between the input image and audio signals and is built to handle sequential data. The CNN-performance LSTM's can be improved by adjusting the number of LSTM layers and the number of neurons in each layer, which are both hyperparameters.

Tasks like regularization, dropout, and activation functions might potentially be included in the CNN-LSTM architecture module. By including a penalty term in the loss function, regularization approaches stop overfitting. In order to avoid overfitting, the CNN-LSTM is trained with a technique called dropout that randomly removes some of the neurons. In order to describe complicated interactions between the input and output data, activation functions are employed to bring non-linearity into the CNN-LSTM.

A series of feature vectors representing the likelihood of producing each audio sample is the CNN-output. LSTM's This series of feature vectors is examined by the post-processing module, which then creates the final audio signal.The post-processing module may carry out operations like denoising, spectrogram creation, and inverse Fourier transform in order to do this. Denoising aids in removing any noise that might be present in the feature vector sequence. In order to effectively manipulate and process the audio input, the sequence of feature vectors is represented in the frequency domain using spectrogram creation. The final audio signal is created by converting a series of feature vectors from the frequency domain to the time domain using the inverse Fourier transform.The post-processing module may also perform tasks including spectrum smoothing, dynamic range compression, and waveform normalization. To make sure that the amplitude of the audio signal is within a particular range, waveform normalization is performed. By modifying the audio signal's dynamic range, dynamic range compression can improve its clarity and lower distortion. To lessen any abrupt transitions or abnormalities that might be present in the audio signal, spectral smoothing is applied.

In general, the post-processing module is a crucial component of the CNN-LSTM image to audio conversion system. It examines the CNN-output LSTM's and changes it into a useful illustration of the audio signal. Several methods to improve the final audio signal's quality, such as denoising, spectrogram creation, and spectrum smoothing, may be included in the post-processing module. The post-processing module makes sure that the audio signal produced appropriately reflects the supplied visual data and is of excellent quality.



V. RESULTS

The output of an image to audio conversion using CNN would be an audio waveform, which can be played through a speaker or headphone.The quality and fidelity of the resulting audio waveform will depend on several factors, such as the complexity of the image, the design of the CNN architecture, the size and quality of the training dataset, and the hyperparameters chosen during training. A well-designed CNN architecture with a large and diverse training dataset can produce high-quality audio waveforms that accurately reflect the content of the input image, while a poorly designed or trained CNN may produce distorted or noisy audio outputs.

It is also worth noting that image-to-audio conversion using CNN is still an active research area, and there is ongoing work to improve the accuracy and fidelity of the results.The output of an image to audio conversion using CNN can vary depending on the type of image being used as input. For example, an image of a musical score may produce an audio waveform that represents the musical notes and rhythms, while an image of a landscape may produce an audio waveform that represents the ambient sounds of the environment, such as birdsong or rustling leaves.In addition, the output audio waveform may also be influenced by the settings used during the conversion process, such as the size and resolution of the input image, the number of layers in the CNN architecture, and the choice of activation functions and optimization algorithms. These settings can affect the level of detail and fidelity of the resulting audio waveform.

Overall, the quality of the result for image to audio conversion using CNN can be influenced by multiple factors, including the complexity of the input image, the design and training of the CNN architecture, and the conversion settings used. As this technology continues to advance, we can expect to see more accurate and sophisticated audio outputs from image inputs.

VI. CONCLUSION

In conclusion, this paper presents a deep learning-based approach for converting images to audio descriptions for visually impaired individuals. The proposed system utilizes convolutional neural networks (CNNs) to extract meaningful features from input images and generate corresponding audio descriptions in real-time. The system also incorporates attention mechanisms to focus on the most important features of the images, which enhances the accuracy and relevance of the generated audio descriptions.Overall, this paper contributes to the development of effective and reliable image to audio conversion systems using CNNs, which have the potential to enhance the accessibility and inclusivity of our society by enabling visually impaired individuals to perceive and interact with visual content through audio descriptions. Future work could explore the extension of this approach to more complex image types, such as videos, and investigate the potential for commercial implementation of the system.

REFERENCES

[1] Sneha.C. Madre, S.B. Gundre, “OCR Based Image Text to Speech Conversion Using MATLAB”, Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2019

[2] P Rohit, M S Vinay Prasad, S J Ranganatha Gowda, D R Krishna Raju, Imran Quadri, “Image Recognition Based Smart Aid For Visually Challenged People”, International Conference on Communication and Electronics Systems (ICCES), 2020

[3] Sujata Deshmukh, Praditi Rede, Sheetal Sharma, Sahaana Iyer, “Voice-Enabled Vision For The Visually Disabled”, International Conference on Advances in Computing, Communication, and Control (ICAC3), 2022

[4] Sai Aishwarya Edupuganti, Vijaya Durga Koganti, Cheekati Sri Lakshmi, Ravuri Naveen Kumar, Ramya Paruchuri, “Text and Speech Recognition for Visually Impaired People using Google Vision”, 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021

[5] Abhishek Mathur, Akshada Pathare, Prerna Sharma, Sujata Oak, “AI based Reading System for Blind using OCR”, 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019

[6] Vaibhav V. Mainkar, Tejashree U. Bagayatkar, Siddhesh K. Shetye, Hrushikesh R. Tamhankar, Rahul G. Jadhav, Rahul S. Tendolkar, “Raspberry pi based Intelligent Reader for Visually Impaired Persons”, 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020

[7] Javavrinda Vrindavanam, Raghunandan Srinath, Anisa Fathima, S. Arpitha, Chaitanya S Rao, T. Kavya, “Machine Learning based approach to Image Description for the Visually Impaired”, Asian Conference on Innovation in Technology (ASIANCON), 2021

[8] R. Prabha, M. Razmah, G. Saritha, RM Asha, Senthil G. A, R. Gayathiri, “Vivoice - Reading Assistant for the Blind using OCR and TTS”, International Conference on Computer Communication and Informatics (ICCCI), 2022

[9] S. Durgadevi, K. Thirupurasundari, C. Komathi, S.Mithun Balaji, “Smart Machine Learning System for Blind Assistance”, International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 2021

[10] Sandeep Musale, Vikram Ghiye, “Smart reader for visually impaired”, 2nd International Conference on Inventive Systems and Control (ICISC), 2019