2023 Fall CSE431
Sound of Silence : Transforming Visuals into Audio For the Blind

1st Tahsin Ashrafee Susmit
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
tahsin.ashrafee.susmit@g.bracu.ac.bd

2nd Isratul Hasan
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
isratul.hasan@g.bracu.ac.bd

3rd Maliha Mehejabin
Computer Science and Engineering)
Brac University
Dhaka, Bangladesh
maliha.mehejabin@g.bracu.ac.bd

***Abstract*—This research focuses on developing an innovative solution to help visually impaired peoples in navigating their surroundings more effectively and safely in Bengali. It is a camera-based technology at the core of this system that continuously takes photographs in real-time, processing them at a rate of one picture per second. The system identifies newly entering items in the visual field through image processing and captions them in Bengali. Subsequently, it employs the generated captions from the user's immediate environment to notify the user through audio. Furthermore, one of its most important features of this research is to identify and detect several types of vehicles, including trucks, cars, and buses. Upon identification, the system calculates the distance of these objects and provides auditory feedback to the user. If a vehicle is detected at a considerable distance, the system informs the user about its presence. Conversely, if a vehicle is too close, the system emits an urgent alert to warn the user. By providing the visually impaired with a semblance of "sight" through auditory descriptions and warnings, this research aims to improve sense of place and safety for them, enabling greater confidence and autonomous navigation in their daily life.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The silent struggles of the visually impaired are often missed in the grand tapestry of human experience, where bright colors and beautiful scenery dance before our eyes. Eye impairments affect more than 295 million people worldwide, of which 39 million are confined to permanent darkness. It is not just a number, it is a story of strength and determination in the face of a world that often forgets the quiet beauty of their lives. Everyday struggles and unsaid fears are hidden behind these numbers. Getting around in the world becomes a complicated dance of trust and openness. Surprisingly, about 90% of blind people have accidents. The damage is not just physical, it also sounds in the halls of social neglect. Imagine living in a world where every move you make is a risk and every corner is a maze of unknowns. People who are forced into this silent darkness because of accidents or illnesses are in even more trouble. Not only do they lose their sight, but they also lose a way of life. When independence falls apart, relying on others becomes a painful truth. The quiet of being alone gets louder than any beautiful sight they've seen before. This feeling is even stronger for people who do not have a family to support them. Their journey through the darkness is alone, and they rely on help from others that does not always come through. They fight silently against a world that seems to have forgotten about them. So, in this sea of problems, our study, "Sound of Silence: Transforming Visuals into Audio for the Blind," stands out as a bright spot. It is a symphony of voices uniting to disrupt the silence that envelops individuals with visual impairments, it transcends the status of a personal endeavor. Consider briefly a realm in which the visually impaired could perceive subtle auditory stimuli such as the rustling of foliage, the bustling activity of a market, or the gentle lap of waves. Our study is a dedication to that chance. We are going on a trip to process pictures, use Natural Language Processing (NLP) to turn them into stories, and then play them back as soulful Bengali audio.

It is not just about usefulness, it is also about accepting different languages. By offering Bengali audio material, we are letting people who are blind or have low vision experience the world, hear its silent wonders, and be a part of a story that goes beyond their disability. We are not just coming up with an answer when we work with NLP and audio processing, we are also making a connection. Turning pictures into sounds is not the only thing "Sound of Silence" does, it also changes lives one soul-stirring sound at a time. Come with us on this symphonic trip where silence is not the end but the beginning of a beautiful world where everyone can hear it.

II. Literature Review

A study by Anne Dheeraj Chowdary*1, Samudrala Venkata Sai Sritwik Sreekar*2, and Dr. Cruz Antony J*3 was released in April 2023 and suggested a CNN algorithm for the conversion of text image to audio for visually impaired people. Using OCR technology, they can accurately translate over 38 languages and numerical characters into spoken sentences by scanning photos with a text-to-speech application. The voice processing module and the image processing module are the two main project modules. According to the testing results, Resnet has the lowest accuracy value of 66.06% while

the evaluation metric values for various pictures, such as Inception, Alex Net, VGG Net, and Squeeze Net, have the highest precision value of 84.22%Ṡqueeze Net has the highest accuracy score at 81.21% while ResNet has the lowest recall value at 63.86%Ȧt 83.56% SquuzeNet architecture has the highest overall accuracy when compared to other network topologies.

A research executed by Abdul Hady Akash, Fabiha Faiz Mahi, Md. Arafat Hasnan, Arindam Kishor Biswas, A. A. M. Rahat-Bin-Rafique, K.B.M. Tahmiduzzaman, and Sabuj Kumar Tarofdar was published in 2022. The author suggested a deep learning-based algorithm for captioning Bangla photos. In this investigation, a deep learning-based feature extractor named ResNet50 was employed. The proposed model is specifically designed to generate feature vectors from images, and it incorporates an LSTM network to generate textual captions based on these feature vectors. The research utilized two datasets: the BanglaLekhaImageCaptions dataset, consisting of 9000 photographs, and the Flickr dataset, including 8000 images. The datasets were used to train and evaluate the model upon its completion.

Research on a Novel Approach for Blind - Image to Audio Conversion in Regional Language was published on April 14, 2022. It was conducted by B. Hemalatha, B. Karthik, S. Balaji, G. Vijayalakshmi, and Rabindra Nath Shaw. This research presents a novel tool that helps persons with visual impairments understand handwritten or printed material. The OCR algorithm has been used across the system. The accuracy of OCR is higher than that of current algorithms, according to the results.

An improved encoder-decoder model for Bengali images using deep convolutional neural networks was developed by Muhammad Faiyaz Khan, S.M. Sadiq-Ur-Rahman, and Md. Saiful Islam on February 14, 2021. This research introduces a comprehensive picture captioning system that employs a multimodal architecture. It combines a one-dimensional convolutional neural network (CNN) to encode sequence information with a pre-trained ResNet-50 model image encoder to extract region-based visual attributes. We utilized a total of 7154 photos for the purpose of training. A total of 1000 photographs were utilized for validation, while the remaining 1000 images were employed for testing. Furthermore, empirical findings demonstrate that the CNN language model, when integrated with the merge architecture, effectively captures intricate sentence structure details with enhanced linguistic variety and generates captions that are more precise and human-like compared to the conventional LSTM.

III. Methodology

A. Convolutional neural network (CNN):

This section will discuss transfer learning methodologies and studies on deep neural networks. The CNN is the most often utilized neural network class for the processing of visual images. A multi-layered neural network, serving as the core element of CNN, offers solutions primarily for the examination, classification, and recognition of images and videos. The design of CNN, like the neural network structure of the human brain, draws inspiration from the visual cortex. CNN's recent advancements have mostly been attributed to its ability to glean knowledge from vast datasets, such as ImageNet. CNN consists of three primary layers. The three layers in question are the convolutional layer, the pooling layer, and the fully connected layer. The convolutional and pooling layers primarily facilitate the model's learning process, while the complete connection layer is responsible for carrying out the classification task.

$$S(i,j) = (IK)(i,j) = \sum_{m} \sum_{n} I(m,n)K(im,jn)(1)$$

The convolution process is described by the mathematical equation given in Equation (1). The equation utilizes the variables m and n to indicate the dimensions of the kernel, which is a matrix of dimensions m*n. On the other hand, the variables i and j are used to denote the coordinates of the matrix from which the convolution will be computed. The intricate layers are often separated by a pooling layer. The primary goal is to decrease the size of the feature map, hence reducing the computing resources needed to construct the model. Moreover, by the removal of the model's prominent and unchanging characteristics, it effectively trains the model. The maximum and average pooling layers are the most often used pooling techniques, however there are alternative pooling methods available. Each neuron within the connection layer is linked to every other neuron within the preceding layer. The number of fully connected layers in a CNN architecture may vary depending on its topology. The output layer is positioned subsequent to the last fully connected layer. Output distributions in classification studies are now generated using Softmax regression, which involves gathering probability distributions for the output classes.

B. Inception-v3:

The InceptionV3 architecture, developed by Google's research team, has greatly enhanced convolutional neural networks (CNNs) in the field of image recognition. The distinguishing characteristic of this system is its utilization of Inception Modules, which are intricate arrangements that employ parallel convolutional filters of different sizes and levels of complexity. The Inception Modules serve as feature extractors by integrating many filter sizes, including 1x1, 3x3, and 5x5 convolutions, to collect visual information across different scales. Using this approach, the network is able to accurately comprehend worldwide patterns, textures, and intricate details in photographs. InceptionV3 starts with first layers that extract fundamental features from input images using basic operations like as pooling and convolutions. The network incorporates many Inception Modules as it evolves, progressively improving the retrieved features. Each module enhances the feature representation and augments the network's ability to learn hierarchical representations by considering features at different receptive field widths simultaneously. The design employs

many mathematical processes, including as convolutions, pooling (both max and average), and concatenations, to evaluate and combine information at different scales. Due to this amalgamation of operations, the network has acquired the ability to comprehend intricate details and patterns present in images, hence enhancing the accuracy and robustness of image categorization tasks.

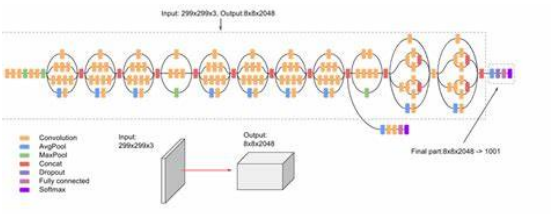

Fig. 1. Inception-v3

C. LSTM (Long Short-Term Memory):

NLP relies on Long Short-Term Memory (LSTM) networks to process and analyze sequential data like sentences and text. In language translation, sentiment analysis, and text production, LSTMs excel in contextual information capture and memory.

NLP uses LSTMs to analyze and understand text sequences. They can detect sentence relationships because their memory cells efficiently handle information flow. An LSTM may preserve sentence context, such as the topic, while processing succeeding words. LSTMs use the Forget Gate to reject unnecessary data and remember contextually relevant data. LSTMs may focus on important linguistic traits and ignore noise or less useful content. The Input Gate of an LSTM also adds fresh data to memory cells. It chooses which new words or linguistic patterns to remember to help grasp the sentence's meaning and context.

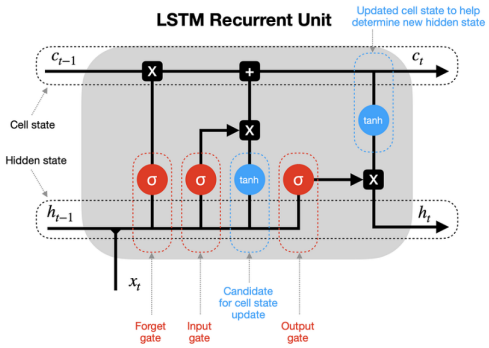

Fig. 2. LSTM (Long Short-Term Memory):

LSTMs can grasp context, subtleties, and links between words in a sequence and make meaningful predictions by integrating these techniques. LSTMs are useful for machine translation, text summarization, sentiment analysis, and language modeling in NLP because they can collect contextual information across phrases or texts.

IV. Dataset

We used the Flickr 30k Dataset, a broad collection of Flickr photographs, to achieve our research goals. The vast 30,000 photos in this dataset provided a broad and diversified foundation for our inquiry. A supplemental dataset curated by Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier helped us correlate written descriptions of images with the visual representations. This extra information illuminated the complex link between language expressions and imagery content. Then translated the English captions of the photographs into Bengali to widen and include our investigation. This translation initiative enabled a more linguistically thorough dataset analysis, especially for Bengali's linguistic nuances and cultural background. We included this translation to guarantee that our findings and analyses were not limited to one language, hence improving their generalizability and applicability.