Sound of Silence: Transforming Visual Data into Audible Information for Visually Impaired Bengali Speakers

Tahsin Ashrafee Susmit, Isratul Hasan, Maliha Mehejabin, Farah Binta Haque, Ehsanur Rahman Rhythm and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
KHA 224, Progati Sarani, Merul Badda, Dhaka - 1212, Bangladesh
{tahsin.ashrafee.susmit, isratul.hasan, maliha.mehejabin, farah.binta.haque, ehsanur.rahman.rhythm}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—This research unveils a groundbreaking solution to revolutionize the navigation experience for the visually impaired Bengali Speaking community. Our approach centers around a robust database of 39,100 images. By harnessing the combined power of VGG16 and LSTM networks, we developed a method to accurately predict descriptions of images in Bengali. Initially, the VGG16 neural network processes the images, extracting key features. These features are then fed into the LSTM network, which generates the corresponding Bengali captions. Through extensive tuning, our model achieved a BLEU-1 score of 0.568526, indicating a significant accuracy in caption prediction. A key aspect of our study is the integration of a specialized tokenizer tailored for the Bengali language, ensuring the captions are not only accurate but also linguistically coherent. To enhance its utility for visually impaired users, we incorporated Google's Text-to-Speech technology to convert these captions into clear, understandable Bangla audio. This study marks a significant step forward in assistive technology, particularly for Bengali speakers with visual impairments, by transforming visual data into audible information, thereby fostering greater independence and environmental awareness.**

I. Introduction

The visually handicapped, a population often disregarded in the diverse range of human encounters, exceed 295 million globally, with 39 million confronting permanent blindness. Aside from statistical data, it is a story of perseverance in a world that occasionally overlooks their understated magnificence. These figures conceal daily challenges and unexpressed anxieties, as maneuvering through the world becomes a subtle interplay of reliance and vulnerability. Surprisingly, about 90% of visually impaired individuals encounter accidents for this negligence. Machine learning is a part of Artificial Intelligence that deals with developing models and algorithms which allows systems to learn new tasks and make decisions autonomously. Deep learning is a subfield of machine learning that relies on a network of several neural networks. Each layer has an input, a hidden layer (or layers), and an output. Each of these levels is composed of interconnected nodes. CNNs, which are a sort of deep neural networks, are specifically designed to handle and examine visual input, including movies and photos.CNN models have achieved notable recognition in several computer vision tasks. The convolutional neural network architecture VGG16 (Visual Geometry Group 16) is renowned for its efficiency and simplicity. It involves convolutional and pooling layers of processing, followed by flattening for the final classification of the image. An additional deep learning model is referred to as RNN. With a hidden state that retains information from previous inputs, a RNN is specifically engineered to process sequential data. LSTM was created to overcome the problem of the vanishing gradient that affects traditional recurrent neural network (RNN) structures. Sequential data long-term dependencies are captured by Long Short-Term Memory networks (LSTMs).

TTS is Google's technology for converting written text to spoken language.It implements neural networks to enhance quality, analyzes input text, models prosody and intonation to simulate natural speech, and provides a variety of voices. Utilizing CNN-VGG16 for image processing, RNN-LSTM for caption generation, and Google Text-to-Speech for audio synthesis, we have developed a solution for the visually impaired as a result of our innovative research. Offering the blind community a valuable resource, this integrated system converts visual data into precise verbal descriptions in a seamless fashion.

Furthermore, a collection of 39,000 captioned images was compiled, and the combined performance of the models yielded a noteworthy accuracy of 97.4082. Significantly, the translation quality is assessed through our Blue scores, which are BLUE-1=0.568526 and BLUE-2=0.387924. Our research is distinguished not only by the combination of various models, but also by our intense concentration on the Bengali language, which imparts an exceptional and virtuous essence to our work. While numerous scholarly investigations examine image-to-caption or text-to-speech in isolation, our paper distinguishes itself by integrating these methodologies, placing particular emphasis on Bengali.

II. Literature Review

In April 2023, Anne Dheeraj Chowdary, Samudrala Venkata, Sai Sritwik Sreekar, and Dr. Cruz Antony J proposed a CNN method to convert text images to audio for visually impaired people. Resnet had the lowest accuracy (66.06%) and recall (63.86%), while SqueezeNet had the highest overall accuracy (83.56%).[1] A 2022 article by Abdul Hady Akash and others presented a deep learning technique for Bangla photo annotation. The study examined 8,000 Flickr and 9,000 BanglaLekhaImageCaptions photographs. [2] On April 14, 2022, B. Hemalatha and others released "Research on a Novel Approach for Blind-Image to Audio Conversion in Regional Language". [3] Muhammad Faiyaz Khan, S.M. Sadiq-Ur-Rahman, and Md. Saiful Islam created a more efficient encoder-decoder model for Ben-gali pictures using deep convolutional neural networks on February 14, 2021. This work introduces multimodal photo captioning. For training, 7154 photographs were used. Validation and testing used 1000 photographs. [4]

In 2016, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun introduced the ResNet model and used the ImageNet dataset. The ResNet-50 variant achieved a top error rate of 6.7%[5] At the time of 2020, Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun likely used benchmark datasets such as PASCAL VOC and COCO to demonstrate superior object detection accuracy, showcasing the model's effectiveness across various real-world scenarios.[6] Gao Huang, Zhuang Liu, and Laurens van der Maaten introduced in the "Densely Connected Convolutional Networks" paper, was tested on CIFAR-10, CIFAR-100 (60k 32x32 images), and ImageNet (1M+ images, 1,000 classes) and it improves information flow, addresses the vanishing gradient problem, and encourages feature reuse, resulting in efficient and effective training. [7] L Abisha Anto Ignatious, S Jeevitha, M Madhur Ambigai, M Hemalatha represented a Semantic Driven CNN-LSTM model for personalized image captioning, utilizing pre-trained CNN, semantic keyword extraction, and facial recognition. Evaluated on Flickr8k and Faces datasets, the model improves captions by integrating semantic labels and personalizing them with recognized celebrities.[8]

He Zhang and Vishwanath Sindagi proposes ID-CGAN, a single image de-raining method using CGANs with a unique generator and multi-scale discriminator. Contributions include improved visual quality and object detection on the VOC dataset using FasterRCNN.[9] Shaik Rafi and Ranjita Das used LSTM on the Flickr 8k dataset with GLoVe embeddings. Inception V3 extracts features, and a Linear Sub-Structure aids caption generation. The model achieves over 81% accuracy, surpassing others, as evaluated by BLEU score.[10] Yossi Adi and Roy Sheffer created image-guided open-domain music generator IM2WAV. IM2WAV makes semantically correct sounds from images or sequences. Finally, IMAGEHEAR, an out-of-domain picture dataset, improves image-to-audio model assessment.[11] Encoding photos to sounds, V. J. Rehna and M. K. Jeya Kumar established an excellent sound steganography encryption technique. The study gives fresh steganography perspectives. Further encryption is done by layering image-converted audio over another music file. The method uses MATLAB 7.0.1 image and signal processing toolboxes.[12]

III. Dataset

Our analysis commenced by initially utilizing an extensive dataset acquired from Flickr 8k. These initial photos played a crucial role in establishing the basis for our inquiry. To enhance the scope and efficacy of our research, we expanded our dataset to incorporate images from Flickr 30k. To expand our dataset, we combined other data sources, resulting in a comprehensive collection of 39,100 photos. In order to promote linguistic variety and enhance accuracy, we have linked each image with a minimum of five Bengali subtitles. By employing a methodical methodology, we effectively conveyed the intricacies of the Bengali language and cultural setting with accuracy. In order to enhance the quality of our dataset, we diligently performed data cleansing, eliminating duplicates and unnecessary captions. In addition, we utilized the 'gtts' (Google Text-to-Speech) module to convert written explanations into an audio version.

TABLE I
Bangla Dataset Visualisation:

Dataset name	Total Sentences	Total Tokens	Total Unique Tokens
Bangla caption	200064	2487896	70436

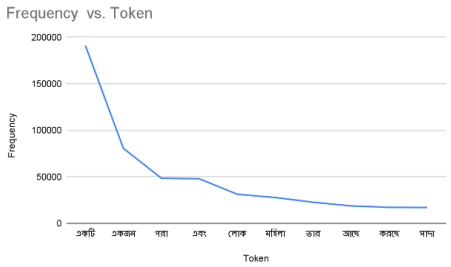

Fig. 1.

IV. 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. Much of CNN's recent progress has been ascribed to its capacity to extract information from large datasets, like ImageNet. There are three main levels in CNN. The convolutional and pooling

layers are mostly in charge of helping the model learn, while the full connection layer handles the classification job.

$$S(i,j) = (I*K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (1)$$

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. Although there are more pooling methods available, the most often utilised pooling strategies are the maximum and average pooling layers. 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. [5][6]

B. LSTM

NLP relies on 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.

It 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.

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, language modeling in NLP because they can collect contextual information across phrases or texts.

C. VGG-16

Finding objects in a picture using 200 different categories is what object localization is all about. The process of image classification entails dividing up all of the photographs into one thousand distinct groupings. In 2014, the VGG 16 model was presented by Andrew Zisserman and Karen Simonyan of the Oxford Visual Geometry Group Lab. In the aforementioned categories, this model was named first and second place at the 2014 ILSVRC.

It requires input picture dimensions of (224, 224, 3). Two levels of 64 channels each, with 3*3 filters and matching padding, make up the top layer. Two convolution layers, one with 128 filters and the other with (3, 3), follow a max pool layer with a stride of (2, 2). After that, a max-pooling layer with the same stride as the preceding one (2, 2) is added. Subsequently, there are two convolutional layers that use 256 filters and(3) filter sizes. Following that, there are two sets of three convolution layers, as well as a max pooling layer. The values of the 512 filters are all (3, 3). The padding is the same for all filters. Two convolutional layers are fed this image. We replace the 11*11 and 7*7 filters used by AlexNet and ZF-Net, respectively, with 3*3 filters in the max-pooling and convolution layers. Input channel count may be changed in certain levels using a 1*1 pixel. To maintain the spatial properties of the image, a 1-pixel padding, called same padding, is applied after every convolution layer.

A map of characteristics was generated by the use of convolutional and max-pooling layers. The compression of this output results in the creation of a feature vector with dimensions of (1, 25088). Three layers that are fully interconnected come next. The first layer produces a vector of dimensions (1, 4096) from the previous characteristic vector. The subsequent layer also produces a vector of size (1, 4096). Nevertheless, the third layer generates a thousand channels specifically for the 1000 ILSVRC challenge classes. To clarify, the third layer that is completely connected applies the softmax function to classify 1000 different categories. Rectified Linear Unit activation is used in every buried layer. ReLU is computationally efficient due to its ability to accelerate learning and mitigate the challenges associated with vanishing gradients.

D. Google TTS

Google Text-to-Speech (TTS) provides a wide variety of authentic voices in numerous languages, dialects, and genders. The purpose of these voices is to imitate the patterns of human speech in order to improve the listening experience by enhancing clarity, intonation, and rhythm. Users have the ability to utilise both male and female voices that come with different accents, therefore providing a diverse range of linguistic options. Advanced vocalisations can also express emotions by modulating pitch, tempo, and timbre. Google TTS offers customisation features that allow users to modify the speech rate and pitch. These options cater to individual preferences and specific applications, enhancing the personalised and adaptable nature of the synthesised speech for different circumstances and user requirements.

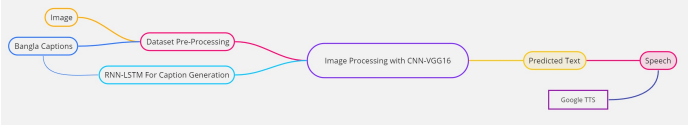

Fig. 2. Work Plan

EXPERIMENTAL SETUP

Both models have been trained using Python 3.11.5 and Tensorflow 2.15.0. The hardware combination comprises an Nvidia GTX 3070 graphics card , 32GB of DDR4 RAM, CPU Ryzen 5600x. While training the models, we utilized a batch size of 32. The dropout rate was defined as 0.1. In addition, RMSprop has been chosen as the optimizer.

V. RESULT

In our research, we analyze the findings derived from the investigation of our project. Furthermore, we analyze and compare two separate models to ascertain their level of accuracy in generating results. During the training phase, we utilised the VGG16 and LSTM architectures to train our models. Through the utilisation of a tokenizer, we have effectively constructed an index of words. This index allows us to make a connection between photographs and captions by means of a mapping mechanism. The training methodology had a batch size of 64 and was executed over 25 epochs. The graph depicting the relationship between epoch and loss ratio is presented in Figure 3.

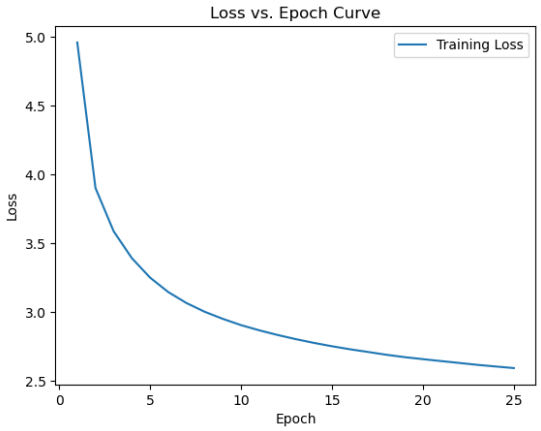

Fig. 3.

The acquired accuracy was 97.4082 % and the reached loss value was 2.5918% . Furthermore, we assessed the performance by employing the BLEU-1 and BLEU-2 scores, resulting in values of 0.568526 and 0.387924, respectively.Furthermore, the utilization of BLEU-1 and BLEU-2 metrics demonstrates the model's proficiency in both individual word matches and sequential phrase alignments, affirming its linguistic precision. Thus, these scores underscore the efficacy of our approach in achieving linguistic fidelity and coherence in the generated content.Throughout the training of the initial model, the average duration for each epoch was 4600 seconds.

The utilization of VGG16 and LSTM architectures, together with mapping methods, tokenizer indexing, and certain training parameters, facilitated both the training process and the evaluation of our model's performance. The findings indicate that the Bangla language excels at generating visual representations that are connected to the text.

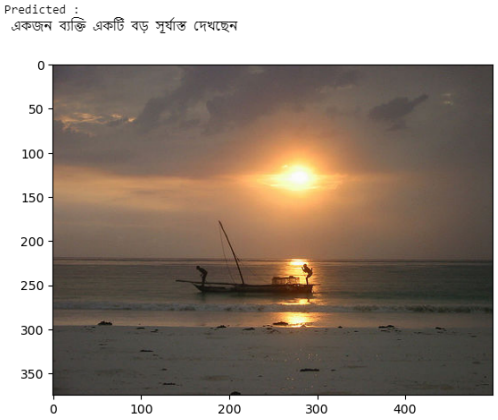

Fig. 4.

Fig. 5.

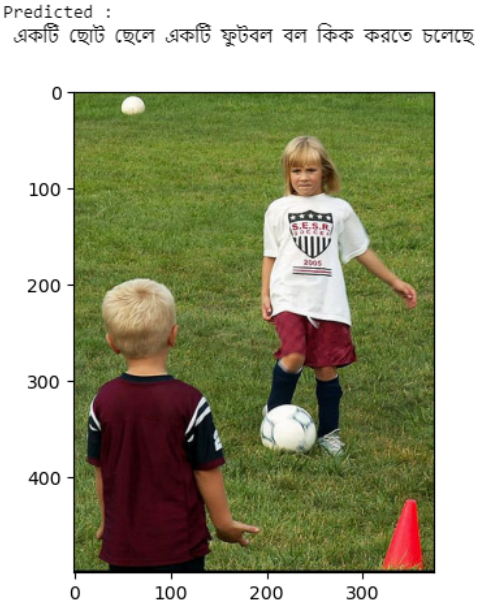

Fig. 6.

In Fig.4, Fig.5, Fig.6 some of the predicted textual repre-

Fig. 7.

sentations are shown while in Fig.7 a random image test in done.

VI. Discussion

The study acknowledges that the suggested method has some inherent flaws while still trying to achieve the ”Sound of Silence”. Natural Language Processing (NLP) picture captioning might be more or less accurate depending on how hard the images are to understand. It might be hard to write captions that fit scenes that aren't clear or are very involved. Also, the pictures that are put into the system have a direct effect on how well it works. Audio quality can be affected when pictures are blurry or don't have enough clarity. This can cause descriptions of the pictures to be less accurate. Then, how well people accept and change the answer determines how well it works. How many people use and agree with the suggested system may rest on their personal preferences, how comfortable they are with technology, and how long it takes them to learn. Moreover, the answer might be useful for a lot of people, but only if it's easy for them to get the technology that it needs, like smartphones and other devices with images and processing power. Thus, the research and development method often runs into these known issues. The most important thing that needs to be done to make the system work better and make the lives of visually impaired people better is to fix those problems.

VII. Conclusion

To conclude, our research brings massive advancement in the assistive technology for the visual impaired Bengali Speaking Community. By the smooth integration of CNN-VGG16 for Image processing along with the RNN-LSTM for caption generation and Google TTS for transforming into audio, we have developed a comprehensive system that has the potential to convert any visual data into Bengali audio descriptions. The remarkable accuracy of BLEU-1 along with the model training accuracy rate, shows the efficiency and reliability of our approach. This approach is often-overlooked and challenges are faced by the visually handicapped, offering them a new dimension of perception and interaction with their surroundings. This research stands as a testament to the potential of AI in enhancing the quality of life for individuals with disabilities, paving the way for more innovations in the future.

VIII. Future Work

The ”Sound of Silence” project is an innovative endeavor that establishes the foundation for future progress in several fields. This innovative initiative centers on many crucial aspects of advancement. The primary objective is to consistently improve algorithms, especially those pertaining to picture description, in order to achieve higher levels of accuracy in understanding visual material. Furthermore, it aims to enhance its multilingual support by using powerful Natural Language Processing models to accommodate a wide range of language subtitles. Furthermore, the research aims to optimize real-time interaction by favoring low latency, which would enhance user experiences in photo processing. Furthermore, it prioritizes user-centric design by utilizing user feedback to develop interfaces that are more accessible and user-friendly. Furthermore, it investigates the incorporation of wearable integration to effortlessly link with wearable technology, resulting in a more streamlined user experience. In addition, the program encourages the use of crowdsourced descriptions, which allows the public to contribute and enhance the system's comprehension of different visual environments. Finally, the initiative encourages cooperation with accessibility groups, establishing alliances to enhance its influence in promoting a more inclusive future for the visually impaired population. The ”Sound of Silence” initiative is dedicated to empowering the visually impaired and promoting a future that is more inclusive and fair by continuously striving for technical advancements.

References

[1] D. Sivaganesan, M. Venkateshwaran and S. P. Dhinesh, ”Image to Audio Conversion to Aid Visually Impaired People by CNN,” 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1707-1713, doi: 10.1109/ICESC57686.2023.10193308.

[2] B. Hemalatha, B. Karthik, S. Balaji, G. Vijayalakshmi \text{and} Rabindra Nath Shaw (2022). A Novel Approach for Blind - Image to Audio Conversion in Regional Language. Springer EBooks, 662–668. https://doi.org/10.1007/978-981-19-1677-9_58

[3] A. H. Akash et al., ”A Deep Learning-Based Approach to Image Captioning in Bengali,” 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), Bhopal, India, 2022, pp. 1-5, doi: 10.1109/CCET56606.2022.10080486.

[4] Faiyaz Khan, M. (n.d.). Improved Bengali Image Captioning via deep convolutional neural network based encoder-decoder model. Retrieved December 10, 2023

[5] He K. Zhang, X. Ren S & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/cvpr.2016.90https://doi.org/10.1109/cvpr.2016.90

[6] Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031

[7] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269. https://doi.org/10.1109/cvpr.2017.243

[8] A Semantic Driven CNN – LSTM Architecture for Personalised Image Caption Generation — IEEE Conference Publication — IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved December 23, 2023, from https://ieeexplore.ieee.org/document/9087299

[9] Image De-Raining Using a Conditional Generative Adversarial Network — IEEE Journals Magazine — IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved December 23, 2023, from https://ieeexplore.ieee.org/document/8727938

[10] A Linear Sub-Structure with Co-Variance Shift for Image Captioning — IEEE Conference Publication — IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved December 23, 2023, from https://ieeexplore.ieee.org/document/9654828

[11] R. Sheffer and Y. Adi, "I Hear Your True Colors: Image Guided Audio Generation," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096023

[12] Rehna, V. J., & Kumar, M. J. (2012). A strong encryption method of sound steganography by encoding an image to audio. International Journal of Information and Electronics Engineering