2023 Fall CSE431 Sound of Silence : Transforming Visuals into Audio in Bengali 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*—In this research, we present an innovative solution aimed at improving the navigation experience for the visually impaired within the 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 0.532, 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, resulting in both bodily scars and a sense of social neglect. Machine learning is a subfield of AI concerned with developing models and algorithms that enable computers 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 97.1391. Significantly, the translation quality is assessed through our Blue scores, which are BLUE-1=0.532062 and BLUE-2=0.353796. 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

A study by Anne Dheeraj Chowdary, Samudrala Venkata Sai Sritwik Sreekar, and Dr. Cruz Antony J 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. The testing findings indicate that Resnet achieved the lowest accuracy rate of 66.06%Ṡqueeze Net achieves the highest accuracy score of 81.21% whereas ResNet exhibits the lowest recall value of 63.86%ṠqueezeNet architecture achieves the highest overall accuracy of 83.56% compared to other network topologies.[1]

The following authors contributed to a 2022 publication: 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. An approach based on deep learning was suggested by the author for the annotation of Bangla images. This research used ResNet50, a feature extractor that is based on deep learning. A long short-term memory (LSTM) network is included into the proposed model to derive textual captions from feature vectors extracted from images. The research used two datasets: one from Flickr (containing 8,000 photographs) and the other from BanglaLekhaImageCaptions (containing 9,000 shots). The datasets were used for both the training and evaluation of the completed model.[2]

On April 14, 2022, the paper ”Research on a Novel Approach for Blind - Image to Audio Conversion in Regional Language” was released. 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.[3]

Muhammad Faiyaz Khan, S.M. Sadiq-Ur-Rahman, and Md. Saiful Islam created a better encoder-decoder model for Bengali pictures using deep convolutional neural networks on February 14, 2021. This study presents a complete photo captioning system based on a multimodal architecture. For training purposes, they used a total of 7154 photographs. A total of 1000 pictures were used for validation, with the remaining 1000 used for testing. Findings from real life also show that the CNN language model, when combined with the merge architecture, does a better job than the traditional LSTM at making accurate and natural-sounding captions that also understand parts of complex sentence structures with more linguistic variation.[4]

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. 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.

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m, j-n)(1)$$

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. 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. [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.

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.

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.

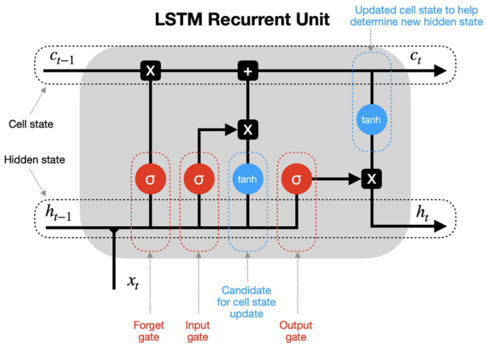


Fig. 1. LSTM:

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.

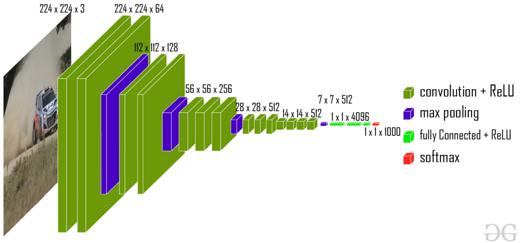


Fig. 2. VGG Architecture

VGG Architecture: 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.

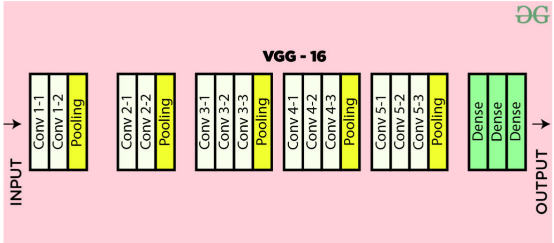


Fig. 3. VGG Architecture map

IV. 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.

V. Result analysis:

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 32 and was executed over 10 epochs. The graph depicting the relationship between epoch and loss ratio is presented in Figure 5.

The acquired accuracy was 97.1391% and the reached loss value was 2.8609%Ḟurthermore, we assessed the performance

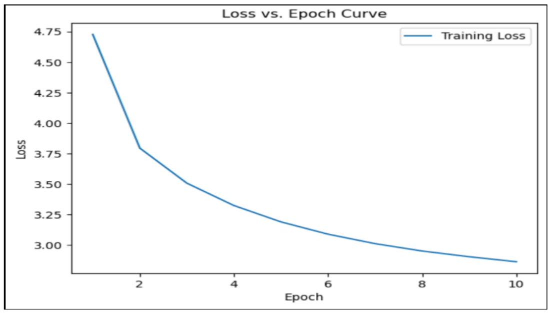


Fig. 4.

by employing the BLEU-1 and BLEU-2 scores, resulting in values of 0.532062 and 0.353796, respectively. Throughout the training of the initial model, the average duration for each epoch was 3789.5 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. Below are some photos together with our models' predicted textual representations of them in Fig.6 and Fig.7.

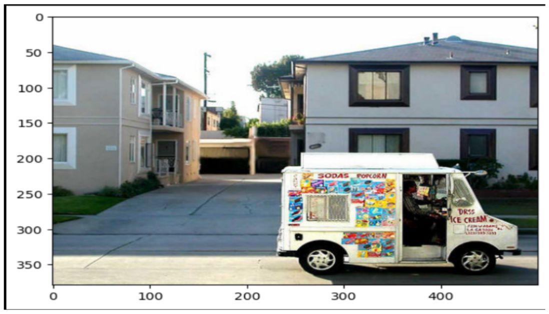
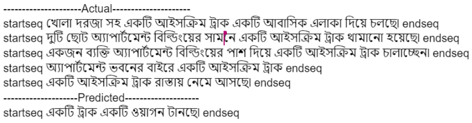


Fig. 5.



```
--------------------Actual--------------------
startseq খোলা দরজা সহ একটি আইসক্রিম ট্রাক একটি আবাসিক এলাকা দিয়ে চলছে। endseq
startseq দুটি ছোট অ্যাপার্টমেন্ট বিল্ডিংয়ের সামনে একটি আইসক্রিম ট্রাক থামানো হয়েছে। endseq
startseq একজন ব্যক্তি অ্যাপার্টমেন্ট বিল্ডিংয়ের পাশ দিয়ে একটি আইসক্রিম ট্রাক চালাচ্ছেন। endseq
startseq অ্যাপার্টমেন্ট ভবনের বাইরে একটি আইসক্রিম ট্রাক endseq
startseq একটি আইসক্রিম ট্রাক রাস্তায় নেমে আসছে। endseq
-------------------Predicted-------------------
startseq একটি ট্রাক একটি ওয়াগন টানছে। endseq
```

VI. Prototype:

The prototype we created functions similarly to a smart storyteller for images. It talks about what's in an image using advanced tech like VGG16 and LSTM. After a lot of practice, it got pretty good, like 97.1391% accurate! That is, it interprets images almost as well as humans do. We can see a prototype example in Fig.8

VII. Google TTS voice:

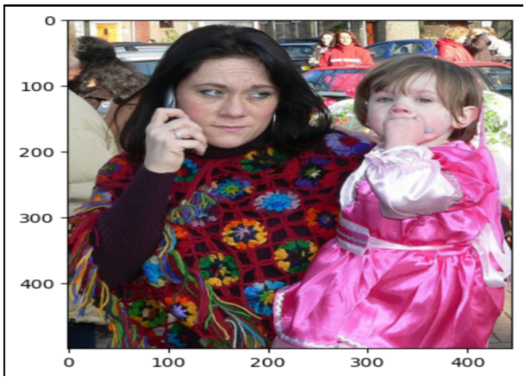
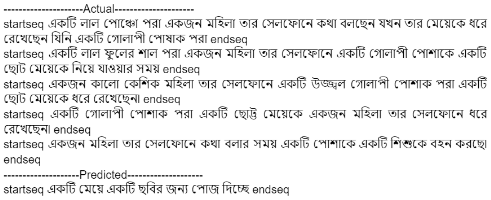


Fig. 6.



```
--------------------Actual--------------------
startseq একটি লাল পোঞ্চো পরা একজন মহিলা তার সেলফোনে কথা বলছেন যখন তার মেয়েকে ধরে
রেখেছেন যিনি একটি গোলাপী পোষাক পরা endseq
startseq একটি লাল ফুলের শাল পরা একজন মহিলা তার সেলফোনে একটি গোলাপী পোশাকে একটি
ছোট মেয়েকে নিয়ে যাওয়ার সময় endseq
startseq একজন কালো কেশিক মহিলা তার সেলফোনে একটি উজ্জ্বল গোলাপী পোশাক পরা একটি
ছোট মেয়েকে ধরে রেখেছেন। endseq
startseq একটি গোলাপী পোশাক পরা একটি ছোট্ট মেয়েকে একজন মহিলা তার সেলফোনে ধরে
রেখেছেন। endseq
startseq একজন মহিলা তার সেলফোনে কথা বলার সময় একটি পোশাকে একটি শিশুকে বহন করছে।
endseq
-------------------Predicted-------------------
startseq একটি মেয়ে একটি ছবির জন্য পোজ দিচ্ছে endseq
```

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.

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.

Limitations

The study acknowledges that the suggested method has some inherent flaws while still trying to achieve the ”Sound of Silence”:

1. Accuracy of Image Captioning: 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.

2. Dependency on Image Quality: The pictures that are put into the system have a direct effect on how well it works.

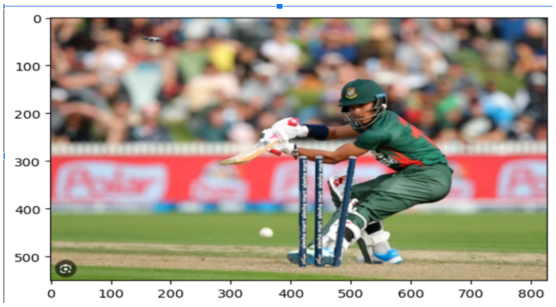
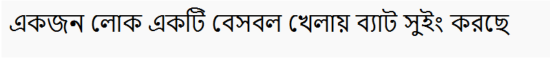


Fig. 7.


একজন লোক একটি বেসবল খেলায় ব্যাট সুইং করছে

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.

3. User Adoption and Adaptation: 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.

4. Technological Accessibility: 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.

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.

Future Work

The ”Sound of Silence” project lays foundations for future progress, such as:

1. Enhanced Algorithms: Image description algorithms are always getting better so that they work more accurately and help people understand what they're seeing.

2. Multilingual Support: Addition of more languages, with strong Natural Language Processing models to handle different language subtitles.

3. Real-Time Interaction: Optimization for picture processing in real time, lowering latency and making the user experience better.

4. User-Centric Design: Using user comments to make things easier to use and more accessible through smart design.

5. Wearable Integration: Investigation of how to easily connect with wearable tech for a lighter option.

6. Crowdsourced Descriptions**: Adding detailed information from the public to help the system understand a wider range of visual situations.

7. Collaboration with Accessibility Organizations: forming partnerships with these groups to make the system's effects bigger.

The ”Sound of Silence” project wants to keep giving the visually disabled community more power by pushing the limits of technology for a better and more inclusive future.

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] Guo, T. (2017). 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA).

[6] Lin, M. Chen, Q. Yan, S. (2013). Network In Network. ArXiv.org. https://arxiv.org/abs/1312.4400