Blind assistant tool using deep learning

[1]Gopal Reddy K A, [2]Hamsa M N, [3]Harish B,[4]Rakshitha V, [5]Dr.Jayavrinda vrindavanam

[1234]B.Tech
Dept. Computer Science and Engineering
Dayananda Sagar University
Banglore, India.

***Abstract*—This According to the World Health Organization, around 40 million people are blind, and another 250 million have some sort of vision impairment. Unfortunately, those with vision impairments typically cannot view images. Our research proposes an approach that can automatically generate audio description of images, which can substantially assist the visually impaired. There are less opportunities for visually impaired people to interact in today's environment, this paper is necessary. For those who are blind or visually impaired, using an image describer can be an enabler. The processed visuals that the visually handicapped cannot see are then given a suitable description and outputted as voice. A simplistic web-interface is developed using Stream lit. The paper uses the Inceptionv3 model as the feature extractor. The Transformer encoder-decoder model to extract image characteristics and generate the text description of the image, which is then converted to an audio using Google Text-to-Speech converter. The generated captions must now precisely reflect the image's graphical information and be highly syntactically understandable. BLEU, METEOR, ROUGE-L, CIDEr score is used for evaluation after captioning. The results were shown to be more accurate, and as a result, they may help the blind access digital media.**

***Index Terms*—Text description, Text-to-speech converter, Inceptionv3, Transformer encoder-decoder model, BLEU**

I. INTRODUCTION

Loss of eyesight or the inability to see the surroundings are referred to as visual impairment. The use of text messages, images, and videos in social media has increased in society. As a result, there may be increased access restrictions unless these media are accessible friendly to the visually impaired. People experience numerous problems on a regular basis, particularly while travelling from one location to another on their own. People frequently rely on assistance from others to meet their daily needs. A number of technologies have been created to help the blind and visually impaired, and one solution to make life easier for the visually impaired has been to develop digital tools that would enable the visually impaired segments to be included in mainstream society. In order to enable the visually handicapped to access visual digital media, this article suggests adding to the existing technologies that help them communicate more effectively by creating audio subtitles for the images.

Many attempts have been made to generate audio subtitles of the photos. One such attempt is that we would wish to make a Deep Learning based System that allows the blind victims to identify real-time objects generating voice feedback .We developed a Web interface using Streamlit library. The user can capture live image using camera or upload image file as its input and focuses on feature extraction using Inceptionv3 model. A Transformer-based encoder creates a new representation of the inputs after receiving the extracted picture characteristics from the Transformer Encoder model. The TransformerDecoder model attempts to learn to produce the caption using the encoder outputs and the text data (sequences) as inputs. A meaningful description of the image is then transformed to an audio output using GTTS API (Google Text to Speech Application Programming Interface). One can distinguish between good and terrible generated captions using BLEU, METEOR, ROUGE-L, CIDEr scores. The results were shown to be more accurate which help blind people feel comfortable around him. The goal of this study is to increase understanding of the relationship between visually impaired readers' information needs and how they engage with various types of images (photographs, drawings, and infographics).

II. REVIEW OF LITERATURE

[1] Deepthi Jain B, Shwetha M Thakur and K V Suresh "Visual Assistance for Blind using Image Processing" April 3-5, 2018 This research describes a cutting-edge method for helping persons who are blind. The suggested system's straightforward architecture and user-friendliness enable the subject to be independent in his or her own house. The technology also seeks to assist the blind in navigating their environment by spotting obstacles, finding their essentials, and reading signs and texts. Initial tests have produced encouraging results, allowing the user to safely and freely move about his environment. By allowing speech to be used as an input to access his basic needs, the system is made considerably more user-friendly.

[2] Jayavrinda Vrindavanam, Raghunandan Srinath , Anisa Fathima ,S.Arpitha, Chaitanya S Rao , T. Kavya "Machine Learning based approach to Image Description for the Visually Impaired", 2021 The alternative method presented in this paper may automatically produce audio descriptions of images, which can significantly help the visually impaired. It uses the Bahdanau attention model and the Inception Resnet - V2 model (MSCOCO dataset) as the feature extractor and decoder (GRU-RNN) to create a text description of the image, which is then converted to an audio file using the Google Text-to- Speech converter (GTTS API).

[3]"End-to-End Transformer Based Model for Image Captioning" YiyuWang1,Jungang Xu2*, Yingfei Sun1 This work presents a pure Transformer-based model that achieves end to-end training and incorporates picture captioning into a single step. When attempting to extract grid-level features from provided images, they adopted Swin Transformer in favour of Faster R-CNN as the backbone encoder. The decoder decodes the refined features into captions word by word. The refining encoder refines the grid features

by capturing the intrarelationship between them. Moreover, mean pooling of grid data is calculated to improve the interaction between multimodal (language and visual) features to improve modelling capability.

[4] Raju Shrestha "A transformer-based deep learning model for evaluation of accessibility of image descriptions" February 2022. In this research, a unique transformer based deep learning model, which can automatically evaluate the quality of a given image description in terms of compliance to the 10 NCAM image accessibility principles, is proposed. By generating modest negative values rather than zero values, RELU(activation function) aids the network in guiding its weights and biases in the desired directions. The models' performance is assessed using the Precision and Recall metrics since they work well with unbalanced datasets.

[5] Imdad Ali Dar,Mayur Anil Bopche, Shubhamsen Mandar Halde, Prof..Rajesh A. Patil "GENERATION OF CAPTION FROM IMAGE AND TEXT-TO-SPEECH CONVERTOR" A web programme with a GUI leveraging Python and machine learning methods. RNN and LSTM are combined with the emerging model in this model to provide photo captions. It creates the most popular captions from the Flickr-8k dataset and converts them into voice for visually impaired users using the TTS engine.

III. MODEL ARCHITECTURE

Datasets

For this project we used Microsoft Common Objects in Context, or MS COCO dataset(2017), where there are several photos in this collection with more than 90 distinct object types. Every year, this dataset is updated, and with each consecutive update, the quantity of photographs likewise increases. Over 1,18,000 photos are used for training, 5,000 for validation, and 41,000 for testing in the most recent dataset version.

Design

For this project, the COCO (Common Objects in Context) dataset will be used to train a deep learning model that will produce written description of the photos. The project's objective is to create a model that can produce natural language captions for a specific image. This model will be used for a variety of purposes to assist visually impaired people, enhance image retrieval systems, and provide descriptions for images on social media platforms.

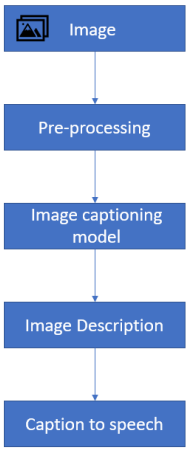

Fig. 1. Model Design

Pre-processing

Data cleansing is the first stage since we need to ensure that the data is in the proper format to work with our model.

- All uppercase characters are being changed to lowercase characters.
- Any character that is followed by a word character ()and a whitespace character () matches one or more consecutive whitespace characters.
- Remove any gaps that may be present at the beginning or the conclusion of the phrase.
- Add sentence start and end tags.

Pre-processing for captions.

- We want the decoder to know when to start and stop the decoding process of a sentence, we format sentences as follows: start + [caption right here] + end. The captions must be padded with tokens to the same length because we distribute them as fixed-size Tensors and their lengths fluctuate naturally.
- Processing format: The model must be given captions in the form of an Int tensor with dimensions N and L, where L is the padded length.
- Tokenization: We developed a dictionary that associates each distinct term with a numerical index. As a result, this dictionary's entry for each word we encounter will have a matching numeric value.

\

Inception V3 Model

InceptionV3's fundamental design is based on GoogleNet The utilisation of Lin's "Network in Network" approach, which increased the representational capability of neural networks is one of the essential components of the Inception architecture. As a result, the dimension was reduced to 11 convolutions, lowering the calculation cost. The Inception architecture was created to lower the computational expense of deep learning-based image categorization . There are typically three possible convolution sizes and one maximum pooling in the Inception module. The basic architecture of InceptionV3 constitutes of the following inception modules: A dropout layer, a fully connected layer with 1024 neuron units and ReLU, and an average pooling layer with a filter size of 5x 5 and stride 3 are all used for dimension reduction. The channel is aggregated following the convolutional operation, and the fusion operator is then applied to the output of the preceding layer. As a result, it aids in lowering overfitting and enhancing the network's flexibility. As the goal of this network is picture captioning rather than classification, we did away with the Softmax layer for classification in our approach to InceptionV3. Each inception module's goal is to record features at various levels. 5×5 convolutional layers are used to capture high-level features, 3×3 convolutional layers are used to capture distributed data, and max pooling layers are used to extract low-level features.

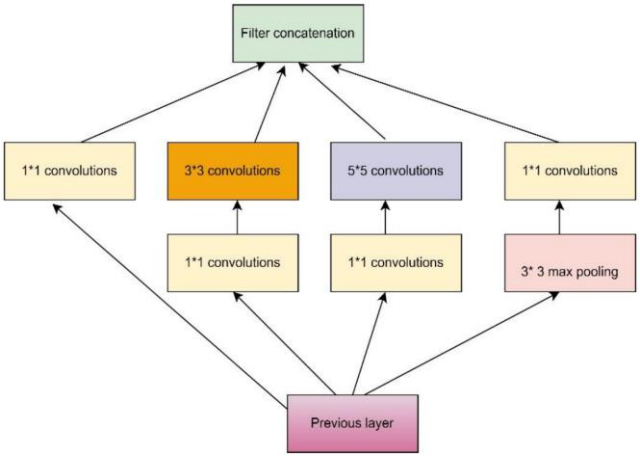

Fig. 2. The architecture of InceptionV3

The image description recieved from the model is evaluated using various evaluation metrics like BLEU score, ROUGE L, METEOR, CIDEr. We got BLEU score for one gram around 0.7961.If BLUE score value is 1 then the model is 100 percent accurate. In our project we are running 20 epochs which results validation accuracy of nearly 0.4258. A graph is plotted for training loss against validation loss. All the evaluation

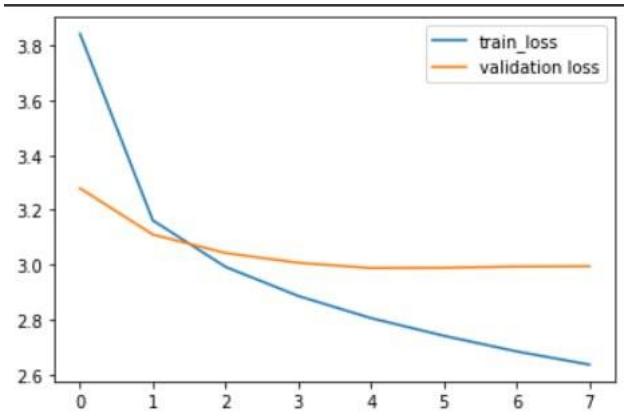

Fig. 3. Transformer model

criteria are evaluated based on the hypothesis and reference sentence. Compute the matrix based on the ref list and hyp list[]. Consider hyp list and search for the word x in the given sentence. The process jumps to the ref sentence and the model will search for the word x in the given sentence. Thenmindividual computation happens the process is shown below.

res = computeindividualmetrics([ref1[0]]; hyp[0])
res = computeindividualmetrics([ref1[1]]; hyp[1])
hyplist = hyp
reflist = [ref]
res = computemetrics(reflist; hyplist)

Transformers

A Transformer encoder creates a new representation of the inputs after receiving the extracted picture characteristics from the Transformer Encoder model. Then Decoder model attempts to learn to produce the caption using the encoder outputs and the text data (sequences) as inputs. Several essential elements are commonly included in the implementation of transformers the input sequence is first translated to a series of embeddings, which are learned during training. Multi-head attention is generally used to

implement the self-attention mechanism, which allows the model to pay attention to several separate input sequence ”heads” at once. The output is then sent via a number of feedforward layers after the self-attention layer, which can aid in capturing more complex interactions between various input sequence components. In every layer, normalization is frequently applied to assist stable the training process and enhance the model’s performance. Transformers take information about the relative positions of various items in the input sequence because they don’t employ any type of sequential processing. Positional encoding is frequently used for this, adding a learned encoding to each location in the input sequence.

The Transformer decoder is made up of a stack of identical layers, just like the Transformer encoder. The Transformer decoder ,however, uses a third major sub-layer for a total of three multi-head attention blocks: The queries, keys, and values are inputs into the multi-head attention mechanism found in the first sub-layer. the multi-head attention mechanism is included in the second sub-layer. And A fully linked feed-forward network makes up the third sublayer. Layer normalization is the step that comes after each of these three sub-layers, where the input to the layer normalization step is its corresponding sub-layer input (through a residual connection) and output. After all these above mentioned process then the text is converted into speech, where in this project we have connected our backend to the web interface where the visually impaired can access our web app easily without anyn complication. In the web interface one can also access the camera and take the picture. For the web interface we have used the Streamlit library[]. The interface is so user friendly where we have included the uploading image feature also and in one click the person is able to take the picture also. The captured image is converted in speech using the NLP technologies and also provieded with selecting different languages for the speech.

WebApp Interface

Streamlit library, an open-source software framework written in Python, is used to create a web interface. A picture can be taken by the interface with just a simple click and get the predicted caption along with TTS in an autoplay config, so that it can be easily accessed by the blind[Fig.6.], also as one chosen by the user from a list of image files as input. To construct a textual description of the image, Inceptionv3 collects the image’s features and passed through the decoding model by analyzing the weights thereby matching the weights with the dictionary of tokens word .then this word will be predicted by sequence to sequence until it predicts the caption or hit the maximum length of 40. This description is then translated to audio using the Google Text-to-Speech converter API[Fig.8.]. The user is given the option to select the language in which he or she is most comfortable having the text and audio displayed[Fig.9.].The application can also be accessed on mobile.

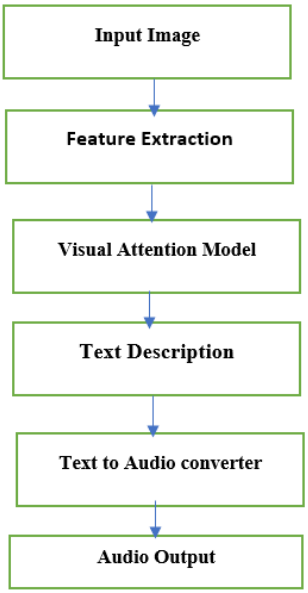

Fig. 4. The conceptual model is as shown in flowchart

IV. EXPERIMENTAL SETUP

Experimental setting

Our objective was to find the optimal set of parameters for each model. We divided the data into two sets in order to accomplishthis: a training set (180000), a testing set (41000), and a validation set (5000). First, we calculated the frequency

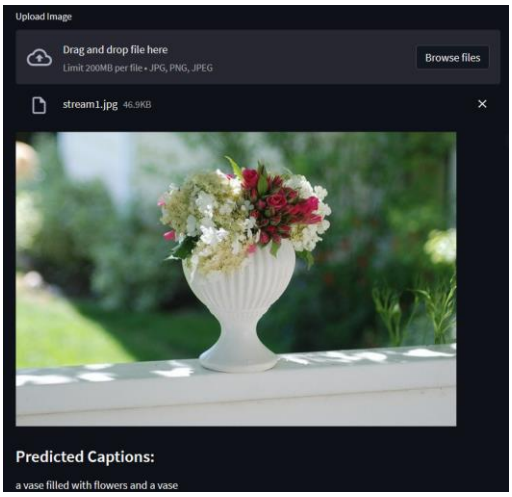

Fig. 5. Web Interface

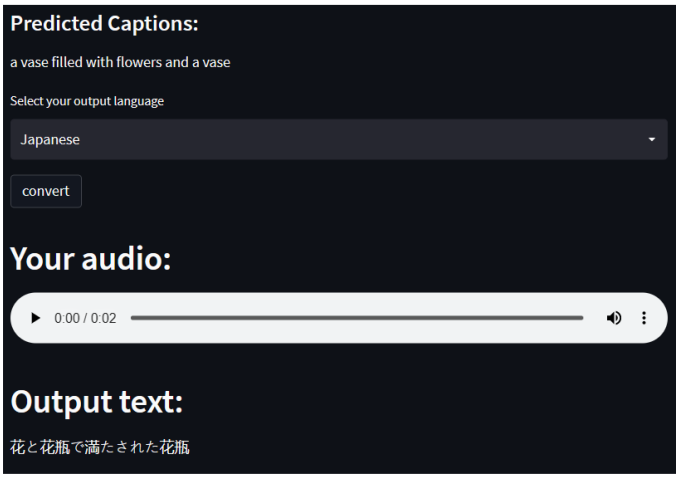

Fig. 6. Web Interface

of the most and least used phrases in our sample. At the preprocessing step, all of the captions were provided and labelled. The goal is for the models to be able to distinguish between the beginning and finish of captions. Using Inception v3, a pre-trained model trained on the dataset, is the next stage. It has fully linked layers and convolutional layers. We did tokenization and built language to provide image captions. Vector notations are created for each word in the caption.

Performance Metrics

We employed a variety of criteria, which are explained below, to evaluate the effectiveness of model prediction : A well-known machine translation statistic that is used to determine how similar two phrases are to one another is the BLEU (Bilingual evaluation understudy) statistic. Papineni et al. made the suggestion. It returns a value; a higher number denotes a closer similarity. You must compare the number of n-grams in one sentence to the n-grams in the reference phrase in order to employ this strategy. Each pair of words is represented by a bigram, whereas a token is represented by a unigram or one gram. The BLEU score of several phrases is calculated using the Corpus BLEU approach. A candidate document isa list where the document is a list of tokens, and a reference list is given using a list of documents.

ROUGE, or Recall Oriented Understudy for Gisting Evaluation, counts the number of ”n-grams” that our model and a reference match produce in the caption. For instance, ROUGEL indicates the use of n-grams, etc.

METEOR (Metric for Evaluation for Translation with Explicit Ordering), in contrast to BLEU, calculates F-score by mapping unigrams.

BLEU SCORE	BLEU_SCORE_VALUES
Bleu_1	0.7961
Bleu_2	0.7097
Bleu_3	0.6395
Bleu_4	0.5819

Fig. 7. Bleu Score Values

EVALUATION CRITERIA	EVALUATED_SCORE
METEOR	0.4361
ROUGE_L	0.7971
CIDEr	5.83908

Fig. 8. Evaluation Criteria

V. RESULTS

After Here are some output results from running the model on the validation part of the MS-COCO database (2017). The photographs are submitted to the system, which then outputs the most popular captions, after being evaluated while keeping in mind the conditioned dataset. Some of the captions that were produced during the first testing and evaluation, together with the corresponding photographs, are supplied below for reference. The system also produces speech from the collected captions.

VI. CONCLUSION

The We have provided a model for attention-based image captioning that can extract text from an image. The suggested model selects the ideal caption for each image and then transforms the selected caption into speech. In order to analyze the model for BLEU score, ROUGE-L, METEOR, and CIDEr, we applied transfer learning on the 2017 MS-COCO dataset. As a result, our model has the atmost accuracy when compared to the other models. Our model has been trained for up to 20epochs, resulting in a loss accuracy of 0.0976. This approach generates voice as the output, which is helpful for people who
are blind.

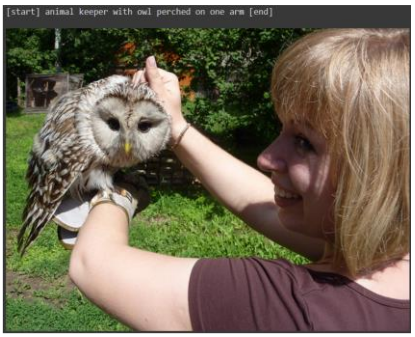

Fig.9. Evaluation Results

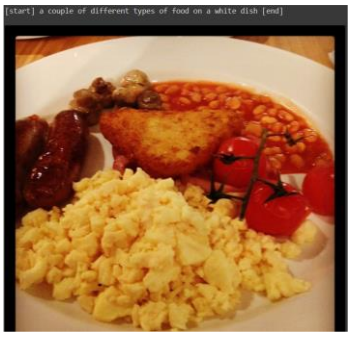

Fig. 10. Evaluation Results

Fig. 11. Evaluation Results

REFERENCES:

[1] Dar, Imad Ali, Mayur Anil Bopche, Shubhamsen Mandar Halde, and Rajesh A. Patil. ”GENERATION OF CAPTION FROM IMAGE AND TEXT-TO-SPEECH CONVERTOR.”

[2] Hoppe, Anett, David Morris, and Ralph Ewerth. ”Evaluation of Automated Image Descriptions for Visually Impaired Students.” In International Conference on Artificial Intelligence in Education, pp. 196-201. Springer, Cham, 2021.

[3] Amritkar, Chetan, and Vaishali Jabade. ”Image caption generation using deep learning technique.” In 2018 fourth international conference on computing communication control and automation (ICCUBEA), pp. 1-

[4]. IEEE, 2018. K. C. Shahira and A. Lijiya, Document Image Classification: Towards Assisting Visually Impaired, TENCON 2019 - 2019 IEEE Region 10 conference (TENCON), Kochi, India, 2019, pp. 852-857, doi:10.1109/TENCON.2019.8929594.

[5] Rajeshvaree Ravindra Karmarkar (2021). Object Detection System for the Blind with Voice Guidance, International Journal of Engineering Applied Sciences and Technology,6(2): 67-70

[6] Jain, B. Deepthi, Shwetha M. Thakur, and K. V. Suresh. ”Visual assistance for blind using image processing.” In 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 0499-0503. IEEE, 2018.

[7] Nganji, J.T., Brayshaw, M. and Tompsett, B., 2013. Describing and assessing image descriptions for visually impaired web users with IDAT. In Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011 (pp. 27-37). Springer, Berlin, Heidelberg.

[8] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In Computer Vision – ECCV 2016, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham,382–398.

[9] Ali Furkan Biten, Lluis Gomez, Marc¸al Rusi˜nol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity- Aware Captioning for News Images. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 12458–12467.https://doi.org/10.1109/CVPR.2019.01275

[10] Himmat Dogra. 2020. A Framework for an automatic evaluation of image description based on an image accessibility guideline. Master’s thesis. OsloMetropolitan University (OsloMet).

[11] S. Amirian and K. Rasheed and T. Taha and H. Arabnia, A Short Review on Image Caption Generation with Deep Learning, the Proceedings of the 2019 International Conference on Image Processing, Computer Vision, Pattern Recognition, 2019, pp.10-18.

[12] P. G. Bhat, D. K. Rout, B. N. Subudhi and T. Veerakumar, Vision sensory substitution to aid the blind in reading and object recognition, 2017 Fourth International Conference on Image Information Processing (ICIIP), Shimla, India, 2017, pp. 1-6, doi: 10.1109/ICIIP.2017.8313754.

[13] B. Makav and V. Kılıc¸, A New Image Captioning Approach for Visually Impaired People, 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 2019, pp. 945-949, doi: 10.23919/ELECO47770.2019.8990630.

[14] BLEU: a Method for Automatic Evaluation of Machine Translation. Kishore Papineni, Salim Roukos, Todd Ward, and Wei JingZhu.