2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)

A Multi-Modal Image Understanding and Audio Description System for the Visually Impaired People

Monisha Preetham
Research Scholar
Monta Vista High School
United States
mk.list@yahoo.com

Meena Krishnan
Technical leader
Cisco
United States
meenakri@cisco.com

***Abstract*— A significant challenge in the field of computer vision is to develop algorithms that are capable of producing captions for images. Ensuring image accessibility is crucial for individuals who have visual impairments. Audio descriptions provide detailed explanations of displayed events, such as bodily movements and expressive gestures in the images. Producing top-notch audio reports necessitates a substantial amount of personal effort in generating descriptions. To overcome this barrier to accessibility, this paper developed a novel multi-modal system that examines the graphical elements of an image and produces audio. The proposed technique extracts the features from the images by applying the transfer learning strategy. The captions undergo pre-processing to generate the tokens. The attention mechanism-encoder-decoder model receives as inputs the tokens and the image features that have been extracted. The model undergoes testing to create captions for the images, and its performance is assessed using the Bilingual Evaluation Understudy (BLEU) score. The average score on the test dataset is 73.864. The produced captions are transformed into audio format to cater to individuals with visual impairments. The real-time implementation of this technology can provide significant benefits to visually impaired individuals, enabling them to comprehend and analyze visual information in real time.**

***Keywords*—** *NLP, Machine learning, computer vision, transformer, CNN, RNN.*

I. Introduction

Amidst the prevalence of visual content, individuals with visual impairments encounter substantial obstacles in accessing and comprehending information conveyed through images. Although there have been advancements in assistive technologies for enhancing accessibility in digital environments, the interpretation of images still poses a significant challenge. Images transmit knowledge across various platforms, including educational resources, reports, online communities, and daily interactions. Individuals lacking visual perception are hindered in comprehending and interacting with the digital realm due to the lack of informative image descriptions [1].

Image captioning is intriguing due to its diverse applications, such as aiding visually impaired individuals, facilitating image indexing, and assisting with other natural language processing (NLP) tasks. Accessing the web through captions for images is essential to blind individuals' daily routines. Simultaneously, visually impaired individuals face significant difficulties identifying images on the internet [2]. Accessing web data and performing everyday tasks such as banking and grocery shopping poses challenges for individuals with visual impairments. The internet is a crucial resource for visually impaired individuals, providing them with significant independence. Therefore, the web accessibility practice provides image descriptions using alternative text (alt text). This alternative text provides concise descriptions as substitutes for the image, conveying the overall message of the image [3].

Including images and captions enables visually impaired individuals to participate in social activities actively, access additional information online, and make informed purchase choices. The computerized creation of captions allows blind individuals to obtain further information about images [4]. Image captioning is the automated process of generating a descriptive caption for an image.

Automatically generating a description in natural language for an image, known as image captioning, poses significant challenges when utilizing a computer. Image captioning necessitates the integration of computer vision and natural language processing fields of research. It entails comprehending the semantic content of an image at a sophisticated level and articulating the information in a sentence that resembles human language. Identifying things' existence, characteristics, and connections in an image is inherently challenging [5]. Arranging a sentence to convey this information further succinctly complicates the task.

Given that a significant portion of human communication relies on written and spoken natural languages, the ability to teach computers to depict the visual world will result in numerous potential uses. These include facilitating natural interactions between humans and robots, enhancing preschool instruction, improving retrieving data, and supporting individuals with visual impairments. Image captioning is a research field in machine learning that is both challenging and significant [6]. It is gaining attention and growing in importance.

Artificial intelligence increasingly emphasizes generating captions for images, which is considered increasingly important. Although AI excels in image recognition, integrating AI with NLP methods is crucial for transforming image captions into speech to assist visually impaired individuals. Natural Language Processing (NLP) facilitates the conversion of written descriptions into verbal speech, offering a hearing interface for individuals with visual impairments. By integrating AI-powered image captioning with NLP, a smooth connection between the visual elements of images and auditory perception can be established, guaranteeing that visually impaired individuals can obtain knowledge in an instructive and captivating way [7].

An important benefit of AI-driven systems for visually impaired individuals is the capability for instantaneous image identification and vocal reproduction. Algorithms powered by artificial intelligence can swiftly process images taken by users' smartphones or other wearable cameras. You can easily access information everywhere with the generated image captions turned into speech. With this real-time functionality's help, people with visual impairments can use and understand their surroundings with a level of freedom and effectiveness previously difficult to achieve [8].

Machine learning (ML) and deep learning (DL) are utilized in image captioning and speech generation. In addition to these applications, ML and DL are employed in the fields of Robotics [9], business [10], databases [11], watermarking [12], and indexing [13]. Thus, this paper presents a novel solution that utilizes advanced deep-learning models to analyze image content and generate meaningful captions. The paper's major contributions are using the pre-trained model, Inception v3, to extract the image features and provide them to the natural language model to generate the text. The customized transformer model, along with the attention module, is proposed to generate the text from the features of the images. The generated text is converted into audio using the customized DL model, which empowers individuals with visual impairments by seamlessly incorporating technology and inclusivity, creating a more inclusive digital environment.

II. Literature Review

Several researchers are transforming images into automated text for descriptive purposes and subsequently converting the text into audio format to cater to individuals with visual impairments. Poongodi et al. [14] suggest a methodology for automatically generating a suitable title and associating a distinct sound with the image. This outcome has been achieved by extensive training and combining two models. Audio suggestions are determined by analyzing the visual content of the situation, while the headings are created using Long short-term memory (LSTM) and convolutional architectures. An accuracy of 67% for the top 5 predictions and an accuracy of 53% for the top 1 prediction have been attained.

Same as above, Kulkarni et al. [15] employ advanced Deep Neural Networks (DNN) such as Convolutional Neural Networks (CNN), LSTM, and methods of transfer learning. The model consists of two distinct stages: 1] Produce descriptive captions for any provided image. Next, the gTTS (Google Text-to-Speech) generator produces audio corresponding to the created captions. This framework is highly advantageous for individuals with visual impairments, enabling them to perceive and understand visual information. The model was trained and tested using the Flickr8K dataset. One thousand more images were used for testing and validation after a total of six thousand images were used to train the model.

Similarly, Sudhakar et al. [16] examine the process of generating image captions using DNN. The model utilizes a combination of CNN, RNN (Recurrent Neural Network), and sentence-generating techniques to generate an English sentence that accurately describes the content depicted in the input image. The caption that is produced is transformed into audio utilizing gTTS technology. These models are constructed using the Flickr 8k dataset comprising over 8000 images.

Alike, Chu et al. [17] introduce an integrated framework called AICRL that utilizes ResNet50 and LSTM with soft attention to generating image captions automatically. AICRL consists of one encoder and one decoder. The encoder uses ResNet50, a CNN, to develop an in-depth representation of the provided image by incorporating it into a series of vectors of fixed length. The decoder utilizes LSTM, a type of RNN, along with a soft attention mechanism. This allows it to selectively concentrate its attention on specific areas of an image to predict the subsequent sentence. The recommended model is utilized for training on a large dataset, specifically MS COCO 2014, to maximize the probability of generating the desired description sentence when provided with the training images.

Likewise, Rahman et al. [18] present the innovative model “Chittron,” a Bangla automatic image captioning system. To resolve the data accessibility problem, 16,000 relevant images from Bangladesh have been gathered and annotated manually in the Bangla language. Model training follows on this dataset to integrate a pre-trained VGG16 image embedding with stacked LSTM layer structures. The model learns to generate captions by predicting one word at a time when given an image as input. The results indicate that the model has effectively acquired a functional language model and has demonstrated high accuracy in generating image captions in numerous instances. Amirian et al. [19] also suggest employing Generative Adversarial (GAN) models to develop novel and different collections through combination. Also, to improve the quality of picture captions, different autoencoders are used.

Xiao et al. [20] also suggested an attentional LSTM model, ALSTM. It is designed to enhance the input vector by leveraging the architecture's hidden states and sequential contextual data. The ALSTM model can prioritize significant features, such as spatial focus on visual relationships, and allocate greater emphasis to the most pertinent contextual words. Furthermore, ALSTM is employed as the decoder in several traditional methods and demonstrates the process of obtaining efficient context attention for changing the input data vector. ALSTM-based approaches can produce explanations of outstanding accuracy by integrating sequencing data and linkages.

Wang et al. [21] proposed a novel system that utilizes a graph neural network to implicitly depict the relationship between important parts in an image, following similar methodologies as mentioned above. In addition, a novel context-aware attention method directs attention selection based on a complete recollection of previously observed visual information.

Also, Sumbul et al. [22] proposed a DL-based image captioning approach. The suggested methodology comprises three primary stages. The initial stage involves acquiring conventional image captions by combining CNN and LSTM networks. Unlike current Remote Sensing image captioning methods, the second step utilizes sequence-to-sequence neural networks to condense the ground-truth captions of every photo used for training into only one caption. This process effectively removes every duplication found in the training set. In the third step, each image's adaptive weights are assigned by default. These weights are then used to combine the conventional descriptions with the condensed captions, considering the image's semantic content. This is achieved by incorporating a novel adaptive weighting technique into the architecture of LSTM networks.

Alternatively, Kuo and Kira [23] examined the graphical model by integrating an extra input to depict absent data, such as object associations. The process entails extracting characteristics and connections from the Visual Genome database and utilizing them to impact the captioning model. A multi-modal pre-trained CLIP model is used to acquire these contextual descriptions. Furthermore, the object detector's outputs remain unaltered due to a frozen model, resulting in a deficiency of the required depth to support the captioning model effectively. Thus, the detector and description outputs are connected, showing that this approach can improve grounding. The efficacy of the suggested methodology is evaluated by producing captions for images.

The literature analysis reveals that most researchers have favored employing CNN, RNN, LSTM, and encoder-decoder models for automated image caption generation, while only a minority have utilized attention models. This paper proposes a novel architecture using the Encoder-Decoder model with an attention mechanism and transfer learning approach to produce textual content based on images. The gTTS API is utilized for audio conversion.

III. Methodology

The paper proposes a novel methodology using the transfer learning approach and developing the customized encoder-decoder model for automated caption generation.

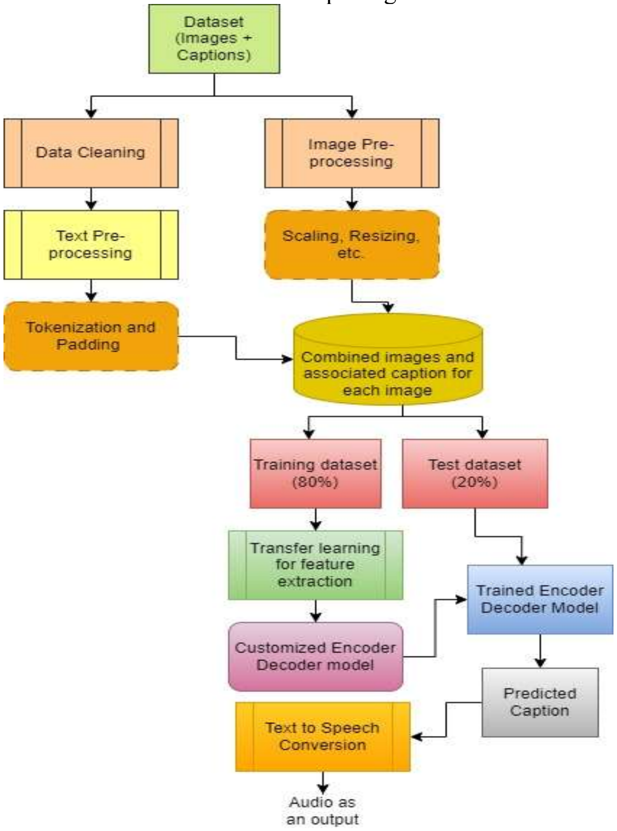
Dataset (Images + Captions)
Data Cleaning
Image Pre-processing
Text Pre-processing
Scaling, Resizing, etc.
Tokenization and Padding
Combined images and associated caption for each image
Training dataset (80%)
Test dataset (20%)
Transfer learning for feature extraction
Trained Encoder Decoder Model
Customized Encoder Decoder model
Predicted Caption
Text to Speech Conversion
Audio as an output

Fig. 1. Proposed Encoder-Decoder-based method and transfer learning for automated caption generation.

The proposed method's design is illustrated in Figure 1. The following explains the methodology that has been suggested.

A. Dataset description

The dataset used for the experimentation is the "Flickr 8k Dataset" [24]. The dataset consists of narrative picture analysis and retrieval, comprising eight thousand photos annotated with five distinct captions that elucidate the key features and occurrences.

A child playing on a rope net

A little girl climbing on red roping

A little girl in pink climbs a rope bridge at the park

A small child grips onto the red ropes at the playground

The small child climbs on a red ropes on a playground

Fig. 2. One of the images, along with captions from the dataset

The photographs were selected from six different Flickr groups. Most pictures do not feature well-known individuals or places; instead, they were chosen by hand to depict various settings and circumstances. One of the images from the dataset and the captions given are shown in Fig. 2.

B. Natural language processing (NLP)

The text data is processed by eliminating punctuation marks and numerical values and converting the text to a consistent case for subsequent analysis. After the data has been cleansed, the text undergoes pre-processing to generate tokens from the text. The process of tokenizing entails breaking a text into smaller pieces. The degree of fineness of the tokenization process determines whether tokens are words, sub-words, or characters. Padding is applied to texts of varying lengths to provide uniformity in the text length for training deep learning models.

C. Image pre-processing

The images are also pre-processed to convert to the same size, as all the images might have different sizes. Thus, all the images are converted into 299 x 299 x 3 as color images.

D. Combining image and text data

The dataset is created by combining the pre-processed text as the target variable and the image as the input variable. The combined dataset is divided into two sets. Eighty percent of the data is used for DL model training, and twenty percent is used for DL model testing.

E. Transfer learning approach

Transfer learning uses pre-trained models for fast training and provides better results with few data samples. The pre-trained Inception v3 model [25] is fed the training dataset to extract features from images.

Inception v3 is a CNN-based architecture created by Google to classify images. Inception modules, known for their deep and efficient design, enable the model to capture features at various scales by incorporating parallel convolutional operations with different kernel sizes. The architecture includes reduction blocks to minimize spatial dimensions and auxiliary classifiers for normalization purposes during training. Batch normalization is implemented over the entire network, which aids in achieving faster convergence. Inception v3 has gained widespread popularity for image-related tasks because it emphasizes effectively utilizing computing resources. The model was initially trained using the ImageNet dataset [26], demonstrating its ability to accurately identify various objects across multiple categories. The 2048 features are extracted for each image

F. Proposed Customized Encoder-Decoder model with an attention mechanism

The encoder is the CNN model, which consists of primary convolution, dense, and pooling layers. The encoder model analyzes the incoming image and produces a standardized representation of a specific size, commonly known as the "image features" or "context vector." The vector acquires the pertinent data from the image and functions as the primary input for the decoder. The resultant feature vector remains constant and does not undergo any changes at each timestamp. Hence, an attention mechanism is required.

The attention method enables the model to concentrate on various regions of the image while generating individual words in the caption. The attention mechanism calculates a weighted total of multiple sections of an image's features during every decoding phase rather than just depending on fixed-size encoded image features.

The decoder is the Gated Recurrent Unit (GRU). The feature vector generated from the attention mechanism is provided as an input to the decoder. The decoder produces a series of words sequentially, one after another. The decoder produces two outputs at each step: the estimated word and a revised hidden state. The updated hidden state is then utilized as input for the following step. Thus, the initial hidden state of the decoder is often determined by using the encoded image features obtained by the attention mechanism, which establishes the context for creating the first word.

G. Text-to-Speech Conversion

The trained model is provided with the test images to predict the caption. Once the caption is predicted, it must be converted into audio so blind people can understand the image's description. The TTS, the Python library that provides text-to-speech conversion, converts the predicted caption into audio.

H. Performance metric

The performance of the proposed model is evaluated using the standard performance metric known as BLEU. It is also known as Bilingual Evaluation Understudy, a metric that assesses the caliber of the text created by machines, particularly in machine translation. The BLEU score quantifies the degree of resemblance between the produced text and one or several standard-given texts. The underlying concept is that a high-quality translation should have similar n-grams (continuous sequences of n elements, such as words) as the standard captions. The equation for the BLEU is given in Eq. 1.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} \omega_n \log(P_n)\right) \qquad (1)$$

Where BP= brevity penalty, BP=1, if the length of the predicted text is greater than or equal to the length, while if it is otherwise, BP is calculated as given in Eq. 2.
N = maximum number of n-grams,
ω_n = Weight given to the precision of N-grams,
P_n = Precision of N-grams, calculated as shown in Eq 3.

$$BP = \exp\left(1 - \frac{len \quad of\ closest\ reference\ text}{leng \quad of\ predicted\ text}\right) \qquad (2)$$

$$P_n = \frac{Number\ of\ matching\ N{-}gra \quad in\ the\ predicted\ text}{Total\ number\ of\ N{-}gra \quad in\ the\ predicted\ text} \qquad (3)$$

IV. Results and Analysis

The experiments are conducted using the Jupyter Notebook on the Kaggle platform. The GPUs used are the Tesla P100. The various libraries of Python, like Tensorflow, gtts, wordcloud, and playsound, are used. The text data is pre-processed to find the top words in the text. The word cloud for the 30 words in the text is shown in Fig. 3. From Fig. 3., it is depicted that the 'in,' 'on,' 'is,' 'and,' 'of,' 'white,' 'girl, 'boy,' etc. are some of the common words in the captions of the images.

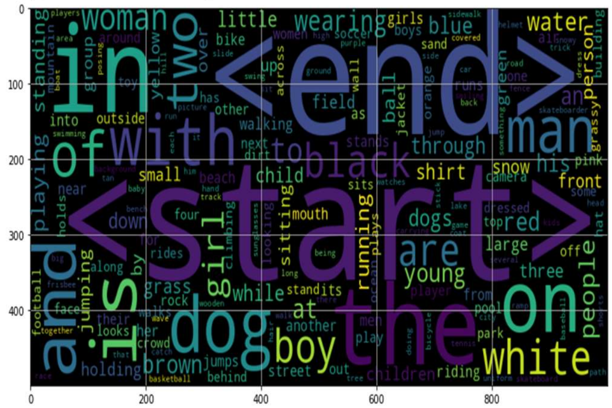

Fig. 3. Word cloud

The pre-processed images in the dataset are given in Fig 4.

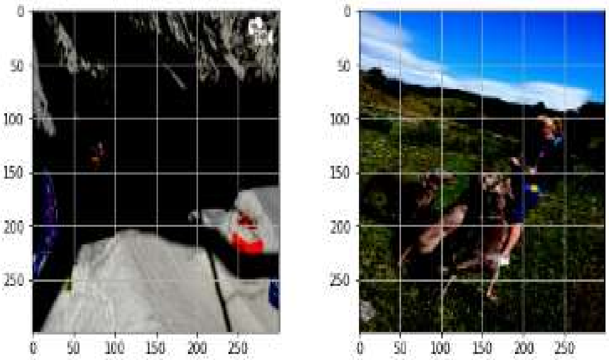

Fig. 4. Pre-processed images

The datasets of pre-processed text and images are merged to create a common dataset. The dataset is now divided into the training set, which consists of 32364 images, while the test dataset consists of 8091 images.

The training data is provided to the Inception v3 model to extract features. These feature vectors are provided to the proposed encoder-decoder model with an attention mechanism. Table I contains the hyperparameters that are appropriate for the customized encoder-decoder model that has been proposed, which includes the attention mechanism.

TABLE I. HYPERPARAMETERS FOR THE PROPOSED MODEL

Hyperparameter	Value
256	256
Embedding units	512
Vocab_size	5000
Batch size	64
Epochs	20
Max_length	31
Dense layers in the decoder	2
Activation function	Rectified Linear Unit (ReLU)
Learning rate	0.01

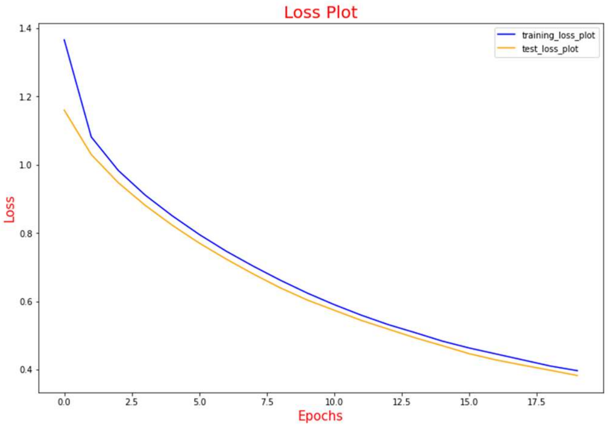

Fig. 5. Comparison of training and test losses v/s epochs.

The training data is provided to the Inception v3 model to extract features. These feature vectors are provided to the proposed encoder-decoder model with an attention mechanism. Table I contains the hyperparameters that are appropriate for the customized encoder-decoder model that has been proposed, which includes the attention mechanism.

The encoder-decoder model, which incorporates the attention mechanism, has been trained, and the graphical representation of its loss is shown in Figure 5. Figure 5 illustrates a significant decrease in loss from the initial to the tenth epoch. During the 10th period, the training loss was 0.624, and the test loss was 0.604. Nevertheless, following the 10th epoch, there is a consistent decline in both the training and test loss. Specifically, the training loss reaches 0.483, and the test loss reaches 0.470 by the 15th epoch. At the 20th epoch, the training loss is 0.397, and the testing loss is 0.382. The time required for training the model is 33.68 minutes on the Tesla P100 GPU.

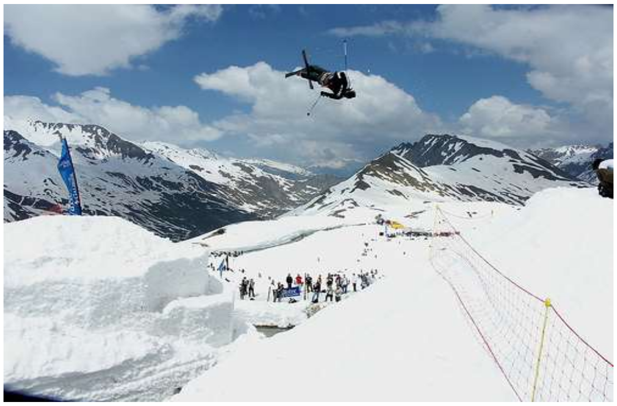

Fig. 6. Example image provided for test to the trained model

The trained model is evaluated using the test data to predict the captions. For the image in Fig 6., the corresponding predicted caption is shown in Fig. 7.

The BELU score for the test image in Fig. 6 is 71.861. The Real Caption is “skier doing flip while jumping,” while the predicted caption is “skier doing flip while the sky.” The audio of the predicted caption is also played for visually impaired people.

The results of the proposed model are given in Table II.

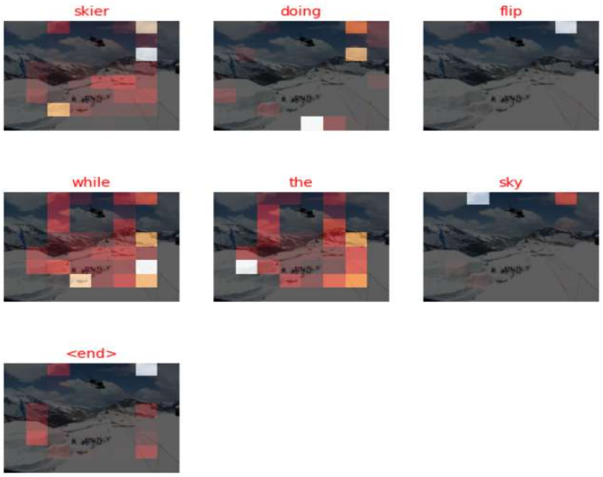

Fig. 7. Predicted Caption

TABLE II. RESULTS FOR THE PROPOSED MODEL

Performance metrics	Value
Train loss	0.397
Test loss	0.382
Training time	33.68 minutes
Average BELU score for test images	73.864

V. CONCLUSION

The combination of AI, speech synthesis, and image caption creation is a powerful tool for removing obstacles for those who are blind or visually impaired. This revolutionary technology enables visually impaired individuals to access content and allows them to navigate the progressive visual digital environment independently and inclusively. The possibility of developing a more accessible and enjoyable browsing experience for the visually impaired is becoming more apparent as the complexities of these AI-driven solutions are explored. Therefore, the research proposes a unique technique that utilizes the inceptionv3, a pre-trained model for obtaining image features. Tokenization is executed to extract the corresponding tokens from the provided captions. The encoder and decoder model with attention mechanism incorporates both the characteristics of the images and the tickets. The model that underwent training produced captions for the test data and was assessed using the BLEU score. The mean score achieved on the test data is 73.864. At last, the generated caption is transformed into audio for visually impaired people. Despite achieving favorable outcomes, the research can be expanded to encompass video analysis and provide video summaries for individuals with visual impairments.

REFERENCES

[1] G. T. Bogdanova and N. G. Noev, “Digitization and Preservation of Digital Resources and Their Accessibility for Blind People,” Advances in Systems Analysis, Software Engineering, and High-Performance Computing, pp. 184–206, Apr. 2019, doi: 10.4018/978-1-5225-7879-6.ch008.

[2] A. Hambley, “Empirical web accessibility evaluation for blind web users,” ACM SIGACCESS Accessibility and Computing, no. 129, pp. 1–5, Jan. 2021, doi: 10.1145/3458055.3458057.

[3] S. Wu, J. Wieland, O. Farivar, and J. Schiller, “Automatic Alt-text,” Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Feb. 2017, Published, doi: 10.1145/2998181.2998364.

[4] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell, “Understanding Blind People’s Experiences with Computer-Generated Captions of Social Media Images,” Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, May 2017, Published, doi: 10.1145/3025453.3025814.

[5] P. Diwakar, “AUTOMATIC IMAGE CAPTIONING USING DEEP LEARNING,” SSRN Electronic Journal, 2021, Published, doi: 10.2139/ssrn.3833851.

[6] S. Sharma and N. Desai, "Data-Driven Customer Segmentation Using Clustering Methods for Business Success," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-7, doi: 10.1109/GCAT59970.2023.10353367.

[7] J. Mun, M. Cho, and B. Han, “Text-Guided Attention Model for Image Captioning,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11237.

[8] K. Tiku, J. Maloo, A. Ramesh, and I. R., “Real-time Conversion of Sign Language to Text and Speech,” 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Jul. 2020, Published, doi: 10.1109/icirca48905.2020.9182877.

[9] S. Patel, D. Israni, and P. Shah, “Path Planning Optimization and Object Placement Through Visual Servoing Technique for Robotics Application,” Journal of Automation, Mobile Robotics and Intelligent Systems, pp. 39–47, Jul. 2019, doi: 10.14313/jamris/1-2020/5.

[10] A. Singh, V. Singh, A. Aggarwal and S. Aggarwal, "Improving Business deliveries using Continuous Integration and Continuous Delivery using Jenkins and an Advanced Version control system for Microservices-based system," 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), Aligarh, India, 2022, pp. 1-4, doi: 10.1109/IMPACT55510.2022.10029149.

[11] R. S. Ghongade, and P. J. Pursani, “Comparison of Relational Database and Object Oriented Database,” International Journal of Modern Trends in Engineering and Research (IJMTER), vol. 1, no. 5, pp. 27-33

[12] D. Israni, and M. Bhatt. "Embedding Color Video Watermark in Image using Orthogonal and Bi-orthogonal Wavelet Transform." In Proc. of International Conference on Advances in Computer Science and Application. 2013.

[13] P. Israni and D. Israni, “An indexing technique for fuzzy object-oriented database using R tree index,” 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp), Dec. 2017, Published, doi: 10.1109/icsoftcomp.2017.8280089.

[14] M. Poongodi, M. Hamdi, and H. Wang, “Image and audio caps: automated captioning of background sounds and images using deep learning,” Multimedia Systems, vol. 29, no. 5, pp. 2951–2959, Feb. 2022, doi: 10.1007/s00530-022-00902-0.

[15] C. Kulkarni, P. Monika, P. B, and S. S, “A novel framework for automatic caption and audio generation,” Materials Today: Proceedings, vol. 65, pp. 3248–3252, 2022, doi: 10.1016/j.matpr.2022.05.380.

[16] J. Sudhakar, V. V. Iyer, and S. T. Sharmila, “Image Caption Generation using Deep Neural Networks,” 2022 International Conference for Advancement in Technology (ICONAT), Jan. 2022, Published, doi: 10.1109/iconat53423.2022.9726074.

[17] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, “Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention,” Wireless Communications and Mobile Computing, vol. 2020, pp. 1–7, Oct. 2020, doi: 10.1155/2020/8909458.

[18] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, “Chittron: An Automatic Bangla Image Captioning System,” Procedia Computer Science, vol. 154, pp. 636–642, 2019, doi: 10.1016/j.procs.2019.06.100.

[19] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, “Image Captioning with Generative Adversarial Network,” 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Dec. 2019, Published, doi: 10.1109/csci49370.2019.00055.

[20] F. Xiao, W. Xue, Y. Shen, and X. Gao, “A New Attention-Based LSTM for Image Captioning,” Neural Processing Letters, vol. 54, no. 4, pp. 3157–3171, Feb. 2022, doi: 10.1007/s11063-022-10759-z.

[21] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, “Learning visual relationship and context-aware attention for image captioning,” Pattern Recognition, vol. 98, p. 107075, Feb. 2020, doi: 10.1016/j.patcog.2019.107075.

[22] G. Sumbul, S. Nayak, and B. Demir, “SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning,” IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 8, pp. 6922–6934, Aug. 2021, doi: 10.1109/tgrs.2020.3031111.

[23] C. Kuo, and Z. Kira. "Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17969-17979. 2022.

[24] “Flickr 8k Dataset,” Kaggle, Apr. 27, 2020. https://www.kaggle.com/datasets/adityajn105/flickr8k (accessed Dec. 22, 2023).

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.

[26] “ImageNet.” https://www.image-net.org/ (accessed Dec. 22, 2023).