English -Malayalam Bilingual Assistive Aid Using Multi Modal Machine Learning Technologies

H.O.Lekshmy[1] **and Swaminathan Jayaraman** [2]

Dept of CSE, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala

E-mail: lekshmyho@am.amrita.edu[1], swaminathanj@am.amrita.edu[2]

Abstract. A larger part of the vision-based assistive device accessible in the market operates in the English language. This limits the scope of the product in multi-linguistic countries. To cope up with this, some assistive devices make use of translation APIs and modules to make an interpretation of English into the regional language, but these translations are not up to the mark when comparing with the original text. In this work, we are proposing an AI-based bilingual assistive device for the visually impaired. We assembled this framework on top of multimodal machine learning techniques and an IoT-based sensory framework. The design incorporates a Pi cam connected with Raspberry pi and headphone as its primary component. Picture captured using the camera is processed using image captioning techniques to generate a textual description. Multi modal machine translation techniques are used to translate this textual information into regional language.

Index Terms Blind Assistive device, Multimodal machine Translation, English -Malayalam Multimodal machine Translation, Image Captioning, Blind Sensory Aid.

1. Introduction

Assistive technology is a trending research topic for the last two decades. Assistive technology mainly focused on developing gadgets, frameworks, equipment's, and software that can improve the lives of people struggling with disabilities like visual impairment, deafness, immobility, and so on [1]. Assistive aides assist individuals with disabilities to live autonomously and to participate in educational institutions and work markets. Disabled people frequently endure negligence and isolation because of their conditions. An effective assistive device can light up their life and open fresh paths for them. This showcases the significance of assistive technology in this digital era. Many research activities were going in this field to develop a proficient assistive innovation for individuals with disabilities. A significant piece of these research activities aimed at developing assistive devices for the visually impaired. As per WHO statistics, at least 2.2 billion people all over the world suffer from visual impairment, which includes 43 million blind people [2]. Forecasters foresee a gigantic expansion in the blind population by 2050. According to their forecasts, the visually impaired populace will increase from 43 million to 115 million by 2050[3]. Over 70 percent of the blind population lives in developing countries like India, China, and Indonesia, Russia and so. India accommodates 20 percent of the world blind populace. Figure 1 shows the top 10 countries with the highest blind population.

Recently, a huge assortment of vision-based assistive gadgets was introduced in themarket. These gadgets incorporate educational aids, navigational aids, travel aids, and so

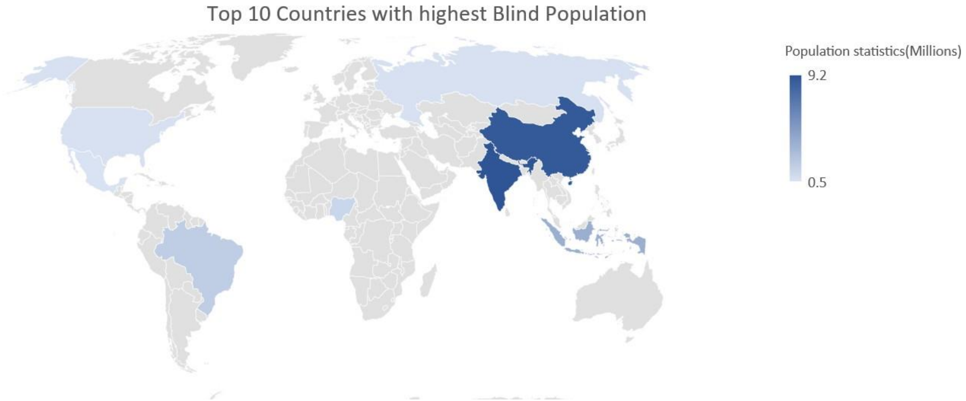

Top 10 Countries with highest Blind Population
Population statistics(Millions)
9.2
0.5

Figure 1: Blinds statistics

on. Most of these products were built on top of computer vision, natural language processing and IoT based Sensory Network. IoT network have a wide range of application which includes development of cost-effective peer sharing models [4], location based alert system [5] and so on Alarge of part vision-based assistive gadgets utilizes English as their operating language. This has been done to maintain universality. but around 70 percent of the world's blind population lives in developing countries, where a fewer percent of the total population speaks English. Customizing an assistive device to a particular linguistic group is not an affordable solution. Thus,80 percent of blind people living in developing countries often face barriers in accessing sensory aids [6].

Development of neutral machine translations [7] partly solves linguistic barriers. Now adays, companies incorporate translation modules that operate on neural machine translations into assistive aids. These modules accept a single context as input and translate it. But these translations are not effective in the case of low resource languages. Most of the time, these translations are literal or word-by-word translations, which eventually lead to misinterpretations [8]. These dangers the existence of visually impaired persons, who utilize computer vision-based sensory aids for navigational purposes. An excellent assistive technology can help blind people with daily tasks, making them more independent and allowing them to achieve a better socioeconomic work–life balance. A device like this could have a significant influence on blind individuals living in poverty. This drives us to develop a useful technologythat can be used by individuals in multilingual countries like India and China. A numerous challenges have arisen on the development stage of this project. The most important of them is lack of high-quality data resources for training.

The major contributions of our work are (i) A schematic design of affordable, effective sensory aid for blind people in the multilingual country (ii)A novel approach to produce high quality translation on low resource languages Indian language like Malayalam (iii)An effective bilingual assistive aid for Malayalam speakers.

2. Literature Survey

Assistive aids for visually impaired people mainly fall into three categories. Electronic travel aids (ETAs), electronic orientation aids, and positional locator devices. Electronic travel aids are the most used blind assistive devices. A large portion of these devices operates on the sensory network and classical computer science algorithms.[9] A breakthrough occurred in this field after the introduction of AI. Many research works have been conducted to integratesAI components with IoT and hardware devices. One such research leads to the development of an AI-Based Visual Aid [10]. This research mainly focused on developing an obstacle avoidance system for both indoor and outdoor environments. It utilizes an integrated IoT framework that fuses a camera, Raspberry Pi, and sensors. Image processing algorithms were used to tackle the object detection problem. Apart from that, it incorporates an OCR-based inbuilt reading assistant for understanding texts. This device was developed for performing simple object detection and there was no arrangement for performing complex tasks like image caption generation.

Afterward, more researchers utilized deep learning models to tackle complex object detection problems. Image captioning is one of such kind. Image captioning aims at generating a textual description of the image using computer vision, natural language processing, and machine language strategies. Three methodologies were created to take care of this issue. The primary methodology is template-based methods. This approach concentrates on detecting objects, actions, scenes, etc., and so on, while the second methodology -Transfer-based caption generation mainly focused on image and caption retrieval techniques. The third methodology purely concentrates on caption generation. Amritkar [11] give prime importance to caption generations using encoder-decoder sequential circuit. This work utilized pre-trained convolutional neural network architecture VGG16 for image feature extraction. Feature vectors and word embeddings were combined using the LSTM model. They tested this methodology on various datasets from numerous dialects.

More advanced versions of the ETa system were later developed. Hengle[12] incorporated more functionalities like image captioning, face recognition, online newspaper reading, etc. into their aid. Another important work in this domain is Dhrishti [13]. It is an outdoor navigational aid make using of computer vison and NLP techniques along with obstacle avoidance. These aids were dedicated to English-speaking users, they had completely ignored native speakers in multilingual countries. At the same time, other sets of aids were introduced that make use of APIs like Google Cloud API for translation. One such approach [14] concentrates on creating an assistive device for blind, deaf, dumb, and blind people with thehelp of Google Cloud APIs. This work utilizes IoT sensory network, Google Cloud Vision API, Text to speech API, and tinker to formulate a sensory aid.

Majority of the translation APIs works on a statistical machine translation technique [15] and classical neural machine translations (NMT) [7]. These machine translations frameworks are not sufficient to provide high-quality translations for low-resource languages. Indian languages are part of low-resource languages as a result it remains as an understudied area. Limited number NLP research were carried out on Indian languages. A few of them where concentrated on Devanagari [16] and Malayalam scripts [17]. This researches mainly concentrate to bring out an optimal solution for character and text identification.

Traditional machine translation techniques produce literal translations, which often leads to misinterpretations. To handle this issue, multimodal AI methods were consolidated into the traditional translation framework. Multimodal machine translation utilizes information gained from more than one modality for efficient machine translation. The most important taskson multimodal translations are speech-guided translation, image-guided translation, and video-guided translation. Multimodal machine translation gives more significant outcomes than normal translation algorithms.[18]. Calixto and Liu [19] made a remarkable contribution to multimodal translation. Their research focused on the exploration of various strategies for integrating global image features into sequential circuits. This research concluded, inserting image features into the decoder is more effective. Later attention-based multimodal translation systems were introduced. Later they have incorporated a double attention mechanism on decoder used for multimodal machine translation [20]. This model is evaluated on European dialects. From 2019 onwards multimodal machine translation become part of WAT (Workshop on Asian Translation). This encourages more research on multimodal translation Rahman [21] used a Multimodal machine translation model builds on top of the NMT tool kit for English -Hindi Translation. This multimodal NMT is trained using a doubly attentive decoder [19]. Later transformers were introduced to perform multimodal machine translation.[22].

3. Proposed model

Proposed system incorporates three modalities they are IoT sensor network, Image Captioning, and Multimodal machine Translation. Figure 2 depicts a schematic depiction of our model.

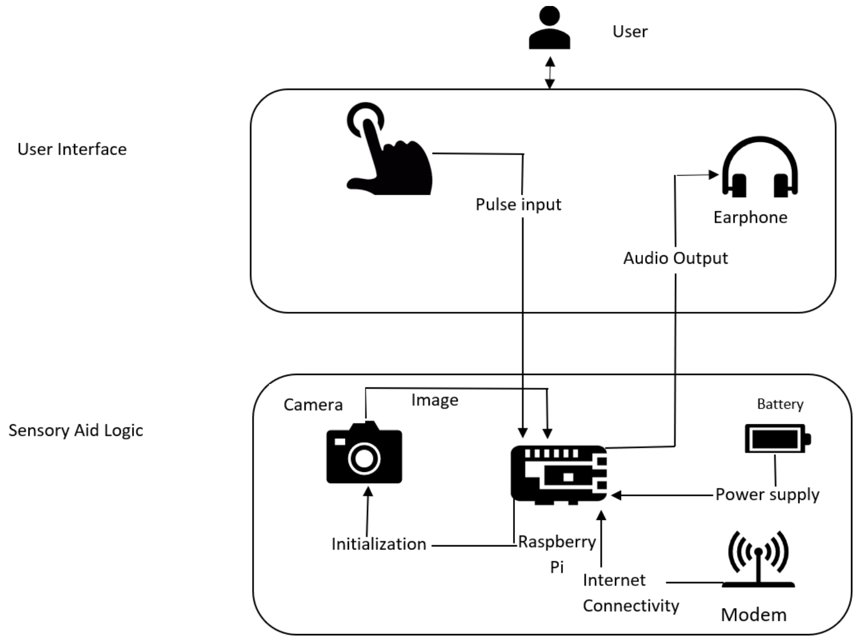


Figure 2: Proposed Model

3.1. IOT Sensor network.

Basic framework of our model is formed by using an IOT sensor network that incorporates Raspberry Pi B3 as its microprocessor, a camera, a push down button, and a headphone.

3.2. Image captioning

Image captioning is a multimodal approach that is commonly used for generating meaningful textual descriptions of images. This process makes use of computer vision algorithms andNatural language processing tools to generate meaningful captions. Image captioning models normally incorporate two sub models. An image model and the language model. Image model comprises a pre-trained convolutional neural network that can be utilized for feature extraction.The language model is made on top a RNN network it is mainly used to generate meaning captions or descriptions from image features.

3.2.1. Feature Extraction We have used a pretrained convolutional neural network called Inception V3 for feature extraction. Inception v3 is pretrained image recognition model that is loaded with weights got for ImageNet dataset. Inception v3 comprises two subparts, a feature extractor, and a classifier, and it have 48 layers which includes multiple fully connected layers. SoftMax layers and corresponding fully connected layers in the Inception architectureare removed for feature extraction. Modified Inception V3 can intake image of size (299,299) and output features vectors of size (8, 8, 2048). So, we have used some preprocessing steps like normalization to match our input images with training model architecture.

3.2.2. Text preprocessing Data cleaning is the first step in text preprocessing, on this step most of the special characters and white space are removed from the captions. Then we had applied tokenizer on captions. Tokenizer splits sentences into tokens or words, a vocabulary is created out of these tokens. Embedding vectors are created with the help of this vocabulary. Padding is used to equalize sentences length.

3.2.3. Model We have generated image captioning system using encoder-decoder circuit with visual attention mechanism. This model is very much similar with Show, Attend and tell [23]. Extracted features are squashed into (64, 2048) and then it is passed through a CNN encoder. This encoder will generate an image encoding (Grated Recurrent units) forms the building block of our language model. We have used Bahdanau Attention [24] in our model. Teacher forcing is also employed on training phase.

3.2.4. Dataset This model is trained on top of World largest Image captioning dataset Coco dataset. Coco dataset comprises 330 k images of 80 classes and 97 orders, which comprises images of different objects which includes peoples, backgrounds, things, animals and so on. This dataset consolidates 5 captions for each image.

3.3. Multi-modal Machine Translation

Multimodal machine translation [25] incorporates one or more contexts to produce effective translations. Multimodal machine translation model architecture is very much like classicalneural machine translation architecture. Both models operate on a sequence-to-sequence circuit that is built on top of the encoder-decoder architecture. The inclusion of extra modality makes multimodal machine translation different from traditional ones. In our case, the image captured by the camera and captions generated by image captioning techniques becomes the input to the multimodal machine translation model. A pictorial depiction of our MMT model is shown in Figure 5

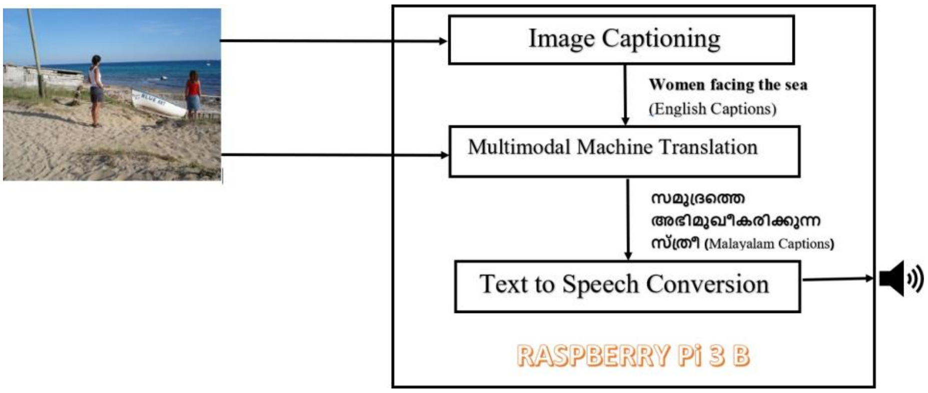


Figure 3: Multimodal machine Translation Model Working

3.3.1. Image preprocessing Multimodal machine translation mechanism incorporates spatial features as an additional input on conventional neural machine translation system. These spatial features are extracted using a pretrained convolutional neural network architecture. In our work we have used VGG 16 (Visual Geometry Group) architecture proposed by Oxford University. VGG 16 is a pretrained convolutional neural network architecture developed for image recognition. This model is initialized with pretrained weights obtained for ImageNet dataset. VGG 16 model incorporates 16 weight layers which includes multiple dense layers. Last layer of VGG 16 is a SoftMax layer which is used for classification purpose. Hence, we had removed last soft max layer from the model. Layers up to maximum pooling layer is used for feature extraction. This model takes an image of input size (224, 224, 3), hence we had converted our input image into a 4D NumPy array of dimension (224 ,224, 3). Generated feature vectors are of the size (,4096). This feature vectors are passed through a dense layer to reshape its size into (,512). As a final step we had pass this vector through a repeat vector. Repeat vector replicates the feature vectors 34 time, which is maximum sentence length of source language.

3.3.2. Text Preprocessing First step text preprocessing is data cleaning. In this part unnecessary characters like symbols, numbers, etc. are removed from sentences in both sourceand target language. We had used ‘start’ as start tag and ‘end’ as an end tag. These tags were appended with each sentence. Then we had split this sentence into words and take count numberof words in each sentence. Figure 3 and 4 depicts Word count representation of English and Malayalam sentences. Then we had applied an in-build tokenizer into our source and target sentence. Tokenizer creates a vocabulary index based on frequency of words and then it will convert all texts in sentence to a sequence of numbers based on this vocabulary. Padding is employed for equalizing sentence length.

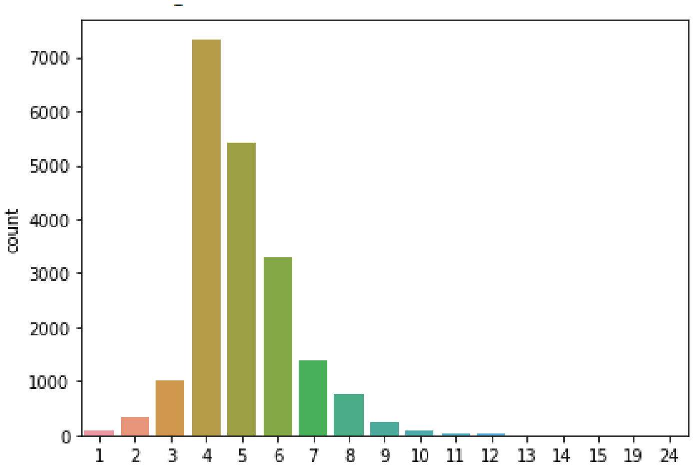


Figure 4: Word count representation of English sentences

3.3.3. Model We have implemented multimodal machine translation system using an Encoder-Decoder sequence circuit build using Keras framework. Our MMT system intakes image and English captions as input and transforms them into translated captions. LSTM (Long short-term memory) cells are used to construct the full encoder decoder circuit. Encoder circuit is constructed using three LSTM (Long short-term memory) cells. An embedding layer is utilizedto convert a source language indexes into a dense vector. Embedding layer forms the first layer in our model. Output from the Embedding layer is passed into the first encoder. We have induced a recurrent drop out of 0.4 and a drop out of 0.4 on each layer. Encoder output from previous states become input for the next encoder cells. Output from the second encoder cell is concatenated with the feature vector. This concatenated vector become input into 3rd encoder LSTM cell. Each encoder gives out an encoder output, hidden state, and cell state. Our decoder networks comprise an embedding layer and a LSTM cell. This embedding layer generates the target language’s embedding vector. The decoder state is initialized using hidden state and cell

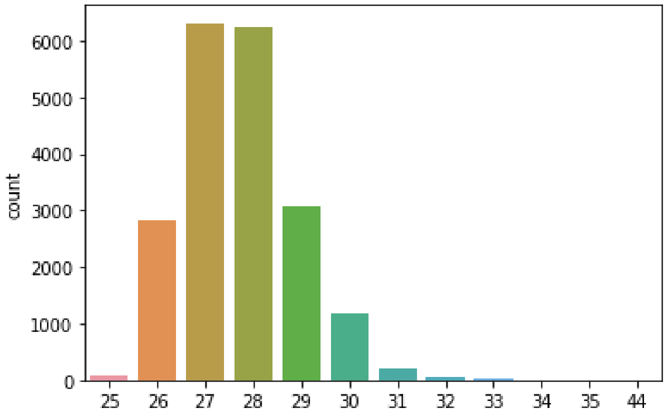


Figure 5: Word count representation of Malayalam sentences

state from encoder. We had used a time distributed layer in the decoder part to wrap the fully connected layers with a single output. Image encodings and embedding vectors are utilized for training process. During the initial phase of inference, we had encoded a test input sequence and retrieved its initial decoder state. Then we had run one step of the decoder with this initial state and a" start of sequence" token as target. For next word prediction, we had set the initial states to the states from the previous time step and continue the above steps. A SoftMax function is used inside the decoder to generate probability distribution over the target vocabulary.

3.3.4. Dataset We have used Malayalam Visual genome Dataset for training our Multimodal machine Translation model. Malayalam Visual genome Dataset set consolidates images, English image captions and its corresponding Malayalam reference test. This data set was realized as a part of shared task in WAT 2021.Malayalam Visual genome Data set comprises of 20 k images for training and 1.6 k for development and 1 k for testing.

4. Overall Working of Model

Our model incorporates a push down button as the triggering switch. Whenever the button gets pushed down, the aid gets triggered, and it will activate the Pi camera connected with Raspberry pi. The surroundings will be captured by this camera. These images are processed using pre-trained models deployed inside Pi. An image captioning model is used to generate meaningful English captions of the image. These captions are later converted into audio waves and delivered throughheadphones. After a couple of minutes device will ask permission for translation process througha voice enquiry. The user can then send voice instructions based on their requirements. We haveused 'yes' as the primary command to grant permission. Aid will intake voice inputs and process the voice inputs using speech recognition API. When the aid detects the permission, it uses a pre-trained Multimodal machine translation model to translate the English captions into Malayalam. This translated output is then transformed into audio and transmitted via a headphone. Workflow of our model is depicted in the Figure 6

5. Experimental Setup

5.1. Hardware Requirements

The following are the hardware components are used in our project

5.1.1. Raspberry Pi 3B The Raspberry Pi 3B is a minicomputer that is widely used in Internet of Things (IoT) projects. This is very similar to a normal computer. Its portability

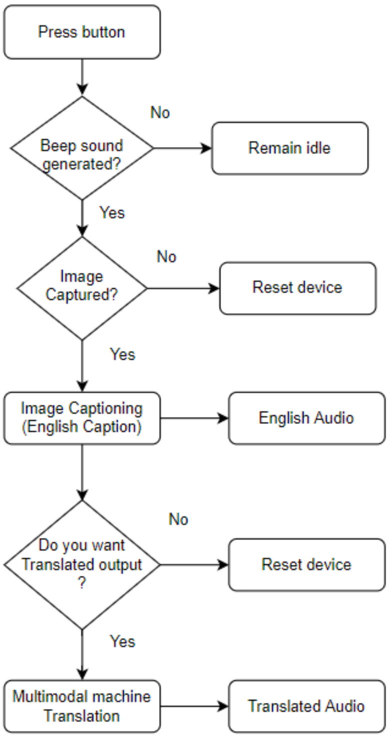


Figure 6: Workflow Diagram

and affordability make it more popular among edge devices. Raspberry Pi provides a Python interface, which makes it is easier for deployment of machine learning algorithms. Different versions of Raspberry pi are available in the market. We are utilizing a Raspberry Pi 3 B with1 GB RAM and a 64-bit processor for this project. Raspberry Pi comprises 4 USB interfacesand camera interfaces, making it suitable for computer vision-based applications.

5.1.2. Camera We are using a Pi camera for this project. Pi cam is a camera made specifically for Raspberry Pi. Raspberry Pi is equipped with a distinct interface for the Pi camera. Raspberry Pi uses MIPI camera serial interface protocol to interact with the camera, It's a 5-megapixel camera module with a resolution of 2592 * 1944 pixels. When compared to other cameras, it is lighter, making it ideal for deployment applications that demand lightweight components.

5.1.3. Push button Here we are using 11x11x4.3MM 4PIN Tactile Tact Push Button as a switch interface to trigger the entire circuit. This switch is act like a reset switch, which can be utilized to reset the entire aid.

5.1.4. Headphone This project uses a Bluetooth-connected headset. Blue tooth headphone is connected with the Raspberry Pi with the help of Pulse Audio package. This headphone can intake audio as an input. It will use packages like pyAudio and speech recognition for audio recognition.

5.2. Software Requirement

We have implemented all models in tensor flow 2.6 framework. Later these modules were converted into tensor flow lite. This helps us to deploy pretrained models on Raspberry pi device without any cloud support. The other major python packages we have used in our project is gtts(Google Text-to-Speech)- gtts(Google Text-to-Speech), it is utilized for text to audio conversion .This library supports more than 50 regional languages which includes Malayalam. Apart from that ,we had used pyaudio and speech recognition packages for voice recognition purposes.

6. Results

6.1. Image Captioning

Image captioning Image captioning models are pre-trained using Coco dataset on GPU K-180 on Colab pro. We have used BLEU score as an evaluation metrics to quantify the quality of generated captions. Our model was able to obtain a BLEU score of 51, which implies high quality scene description. The result obtained for a test case is shown in Figure 5

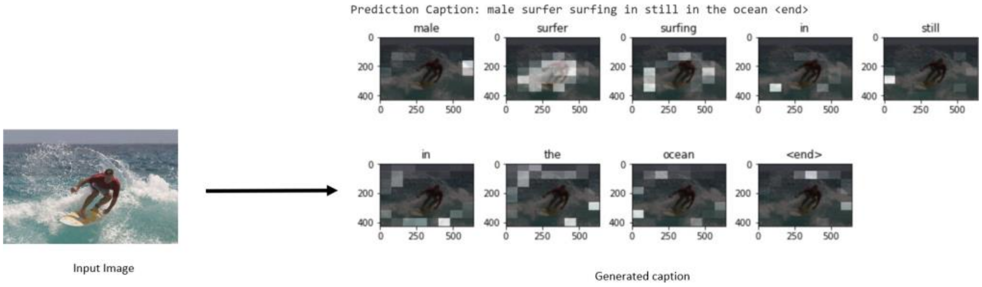


Figure 7: Generated captions

6.2. Multimodal Machine Translation

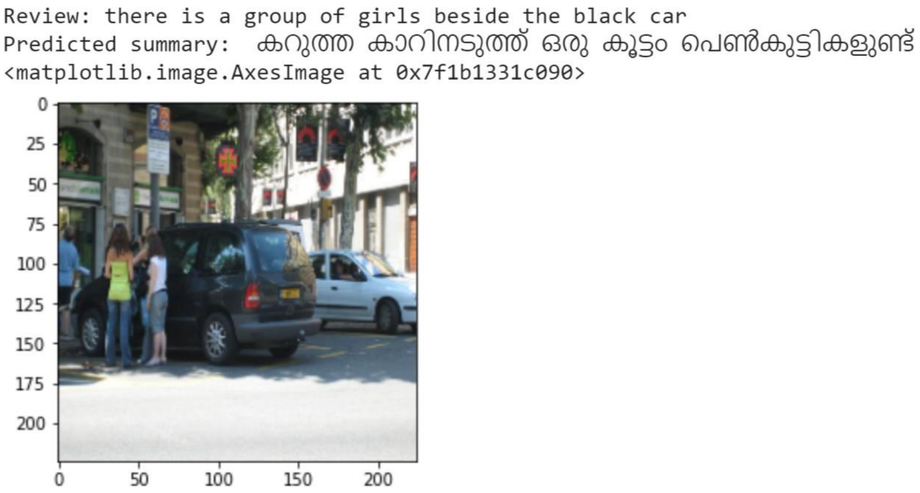
We have trained our Multimodal Machine Translation model on Visual genome Malayalam data. BLEU core is utilized to assess the quality of generated translations. Our model was able to generate an overall BLEU score of 39, which implies understandable translations. A best-case scenario and worst-case scenario of our model is depicted on Figure 7 and Figure 8. In best case, we were able to obtain a BLEU score of 100 and for worst case we got BLEU score 20.

7. Conclusion

We were able to design a sensory that can produce meaningful English captions andits corresponding Malayalam translation. This bilingual aid has utilized multi modal machine learning techniques like Image captioning and multi modal machine translation to achieve astate of art model. This model doesn’t use any cloud platforms for interference as a result, this sensory aid is able to operate in offline Mode. Thus making a affordable solution for Malayalam speaking blind people.

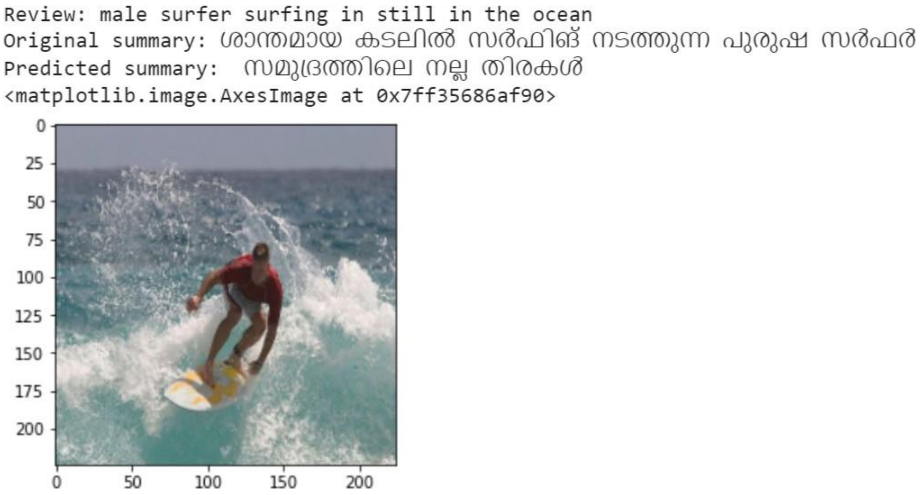
8. Future Work

Translation quality of the Multi modal machine translation system can be further improved using attention mechanism in encoder decoder circuit. A generalized Multi modal machine translation model can be deployed in several practical applications which includes assistive aids, translator aids and so on.



```
Review: there is a group of girls beside the black car
Predicted summary:  കറുത്ത കാറിനടുത്ത് ഒരു കൂട്ടം പെൺകുട്ടികളുണ്ട്
<matplotlib.image.AxesImage at 0x7f1b1331c090>
```

Figure 8: MMT Results-Best Case



```
Review: male surfer surfing in still in the ocean
Original summary: ശാന്തമായ കടലിൽ സർഫിങ് നടത്തുന്ന പുരുഷ സർഫർ
Predicted summary:  സമുദ്രത്തിലെ നല്ല തിരകൾ
<matplotlib.image.AxesImage at 0x7ff35686af90>
```

Figure 9: MMT Results-Worst Case

References

[1] Mohammad O. A. Aqel et al. “Review of Recent Research Trends in Assistive Technologies for Rehabilitation”. In: *2019 International Conference on Promising Electronic Technologies (ICPET)*. 2019, pp. 16–21. DOI: 10.1109/ICPET.2019.00011.

[2] Peter Ackland, Serge Resnikoff, and Rupert Bourne. “World blindness and visual impairment: despite many successes, the problem is growing”. In: *Community Eye Health* 30.100 (2017), p. 71.

[3] Rupert R. A. Bourne et al. “Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis”. In: *Lancet Global Health* 5.9 (Sept. 2017), e888–e897. ISSN: 2214-109X. DOI: 10.1016/S2214-109X(17)30293-0.

[4] H O Lekshmy et al. “An IoT based P2P model for Water Sharing using Machine Learning”. In: *2020 International Conference on Communication and Signal Processing (ICCSP)*. 2020, pp. 0790–0794. DOI: 10.1109/ICCSP48568.2020.9182081.

[5] H Lekshmy et al. “Geo located alert via bottom up IoT model in times of COVID-19 and other disasters”. In: *IOP Conference Series: Materials Science and Engineering* 1128 (Apr. 2021), p. 012001. DOI: 10.1088/1757-899X/1128/1/012001.

[6] Suraj Singh Senjam, Allen Foster, and Covadonga Bascaran. “Barriers to using assistive technology among students with visual disability in schools for the blind in Delhi, India”. In: *Disability and Rehabilitation: Assistive Technology* 3 (Mar. 2020). ISSN: 1748-3107. DOI: 10.1080/17483107.2020.1738566.

[7] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. “A Survey of Multilingual Neural Machine Translation”. In: *ACM Comput. Surv.* 53.5 (Sept. 2020). ISSN: 0360-0300. DOI: 10.1145/3406095.

[8] Ding Liu et al. “A Survey of Low Resource Neural Machine Translation”. In: *2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. 2019, pp. 39–393. DOI: 10.1109/ICMCCE48743.2019.00017.

[9] Alexy Bhowmick and Shyamanta Hazarika. “An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends”. In: *Journal on Multimodal User Interfaces* 11 (Jan. 2017), pp. 1–24. DOI: 10.1007/s12193-016-0235-6.

[10] Muiz Ahmed Khan et al. “An AI-Based Visual Aid With Integrated Reading Assistant for the Completely Blind”. In: *IEEE Transactions on Human-Machine Systems* 50.6 (2020), pp. 507–517. DOI: 10.1109/THMS.2020.3027534.

[11] Chetan Amritkar and Vaishali Jabade. “Image Caption Generation Using Deep Learning Technique”. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. 2018, pp. 1–4. DOI: 10 . 1109 / ICCUBEA . 2018 . 8697360.

[12] Amey Hengle et al. “Smart Cap: A Deep Learning and IoT Based Assistant for the Visually Impaired”. In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. 2020, pp. 1109–1116. DOI: 10.1109/ICSSIT48917.2020.9214140.

[13] A. Kandoth et al. “Dhrishti: A visual aiding system for outdoor environment”. In: cited By 1. 2020, pp. 305–310. DOI: 10 . 1109 / ICCES48766 . 2020 . 09137967.URL: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85091344924 & doi = 10 . 1109 % 2fICCES48766 . 2020 . 09137967 & partnerID = 40 & md5 = 86ee80216e61444c12bd5077c93e35ca.

[14] A. Karmel et al. “IoT based Assistive Device for Deaf, Dumb and Blind People”. In: *Procedia Comput. Sci.* 165 (Jan. 2019), pp. 259–269. ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.01.080.

[15] J. Nair, R. Nithya, and M.K. Vinod Jincy. “Design of a morphological generator for an english to indian languages in a declension rule-based machine translation system”. In: *Lecture Notes in Electrical Engineering* 672 (2020). cited By 0, pp. 247–258. DOI: 10.1007/978-981-15-5558-9_24. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091319077&doi=10.1007%2f978-981-15-5558-9_24&partnerID=40&md5=d0b3a2ca6a5076eb1a421f07d9b5f655.

[16] Vamsi Kikkuri et al. “An Optical Character Recognition Technique for Devanagari Script Using Convolutional Neural Network and Unicode Encoding”. In: Jan. 2021, pp. 173–187. ISBN: 978-981-33-4304-7. DOI: 10.1007/978-981-33-4305-4_14.

[17] S. Thara and P. Poornachandran. “Transformer Based Language Identification for Malayalam-English Code-Mixed Text”. In: *IEEE Access* 9 (2021). cited By 0, pp. 118837–118850. DOI: 10.1109/ACCESS.2021.3104106. URL: https://www.scopus.com/inward/record. uri? eid= 2 - s2 . 0 - 85114317354 & doi= 10 . 1109 % 2fACCESS. 2021 . 3104106 & partnerID=40&md5=cd74f28fe568312d783ec4e9985f4787.

[18] Zhuosheng Zhang et al. “Neural Machine Translation with Universal Visual Representation”. In: *International Conference on Learning Representations*. 2020.

[19] Iacer Calixto and Qun Liu. “Incorporating Global Visual Features into Attention-based Neural Machine Translation.” In: Jan. 2017, pp. 992–1003. DOI: 10.18653/v1/D17-1105.

[20] Iacer Calixto, Qun Liu, and Nick Campbell. “Doubly-Attentive Decoder for Multi- modal Neural Machine Translation”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1913–1924. DOI: 10.18653/v1/P17-1175.

[21] Sahinur Laskar et al. *English to Hindi Multi-modal Neural Machine Translation and Hindi Image Captioning*. Nov. 2019.

[22] Yuqing Song et al. “Enhancing Neural Machine Translation with Dual-Side Multimodal Awareness”. In: *IEEE Transactions on Multimedia* (2021), pp. 1–1. DOI: 10.1109/TMM.2021.3092187.

[23] Kelvin Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: *arXiv* (Feb. 2015). eprint: 1502.03044. URL: https://arxiv.org/abs/1502.03044v3.

[24] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *arXiv* (Sept. 2014). eprint: 1409.0473. URL: https://arxiv.org/abs/1409.0473v7.

[25] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.