AI-Transl:A Multilingual Assistive Device Using Multi Modal Machine Learning

Lekshmy [1,*] **and Dr.Swaminathan** [2]

[1]Dept of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
[2]Dept of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
[*]amenp2ari20020@am.students.amrita.edu
[*]swaminathanj@am.amrita.edu
[+]these authors contributed equally to this work

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 AI-Transl, an AI-based multi-linguistic assistive device for the visually impaired. We assembled this framework on top of multimodal machine learning techniques and an IoT-based sensory system. The design incorporates a Pi cam connected with Raspberry pi and microphone. Picture captured using the camera is processed using image captioning techniques to generate a textual description. Multimodal machine translation techniques are used to translate this textual information into regional language.

Background

Assistive technology is a trending research topic for the last two decades. Assistive technology mainly focused on developing gadgets, frameworks, equipment, 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 the market.These gadgets incorporate educational aids, navigational aids, travel aids, and so on. Most of these products were built on top of computer vision and natural language processing principles. A large 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[4].

Development of neutral machine translations[5] partly solves linguistic barriers. Now a days, companies incorporates 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[6]. This dangers the existence of visually impaired persons, who utilize computer vision-based sensory aids for navigational purposes.

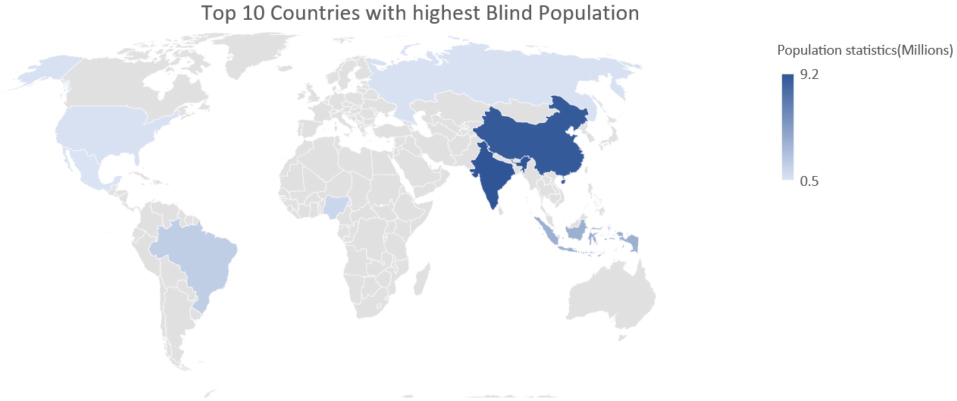
Top 10 Countries with highest Blind Population
Population statistics(Millions)
9.2
0.5

Figure 1. Blinds statistics

Problem Addressed

To develop a multi-linguistic sensory aid for blind individuals, that can produce proficient textual descriptions from images and can change over it into low-asset Indian dialects like Malayalam. This aid is intended to be a mix of 3 distinct technologies,even though the prime focus is given to multimodal machine translation on low resource language. Hence, the underlying problem statement can be addressed in the following manner. To study the impacts of inclusion of additional context as other modalities on neural machine translations and to identify its effects on low resource languages translations, where neural translations regularly fail because of limited training resources.

Related Work

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 commonly used blind assistive devices. A large portion of these devices operates on the sensory network and classical computer science algorithms.[7] A breakthrough occurred in this field after the introduction of AI. Many research works have been conducted to integrates AI components with IoT and hardware devices. One such research leads to the development of An AI-Based Visual Aid[8]. 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[9] 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.This 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[10] incorporated more functionalities like image captioning, face recognition, online newspaper reading, etc into their aid. This aid was dedicated to English-speaking users, on this work they had completely ignored native speakers in multilingual countries. At the same time span, other sets of aids were introduced that make use of APIs like Google Cloud API for translation. One such approach[11]concentrates on creating an assistive device for blind, deaf, dumb, and blind people with the help 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 work on classical neural machine translations (NMT), which operate on sequential encoder-decoder circuits. These Neural machine translations frameworks are not sufficient to provide high-quality translations for low-resource languages. Indian languages are part of low-resource languages. Most of the time, this system produces literal translations, which often leads to misinterpretations. To handle this issue, multimodal AI methods were consolidated into the traditional translations framework. Multimodal machine translation utilizes information gained from more than one modality for efficient machine translation. The most important tasks on multimodal translations are speech-guided translation, image-guided translation, and video-guided translation. Multimodal machine translation gives more significant outcomes than normal translation algorithms.[12]. Calixto and Liu[13] made a remarkable contribution to multimodal translation. Their research focussed 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[14]. This model is evaluated on Europen dialects. From 2019 onwards multimodal machine translation become part of WAT(Workshop on Asian Translation). This encourages more researches on multimodal translation Rahman[15] used a Multimodal machine translation model builds on top of the NMT tool kit for Engish -Hindi Translation. This multimodal NMT is trained using a doubly attentive decoder[13].Later transformers were introduced to perform multimodal machine translation.[16]

Method

Methods. This research is mainly aimed at creating an IoT-based assistive device for blinds using image captioning and multimodal machine translation. Three different functionalities are integrated inside this aid.1 IoT sensor network. 2 Image captioning 3. Multimodal machine translation. Image captured using an IoT network is processed to generate a meaningful English caption. Afterward, this English caption and image become an input for the multimodal machine translation. Multimodal machine translation focuses on translating English captions into other regional languages like Malayalam . An overview of the proposed framework is shown in Figure 2

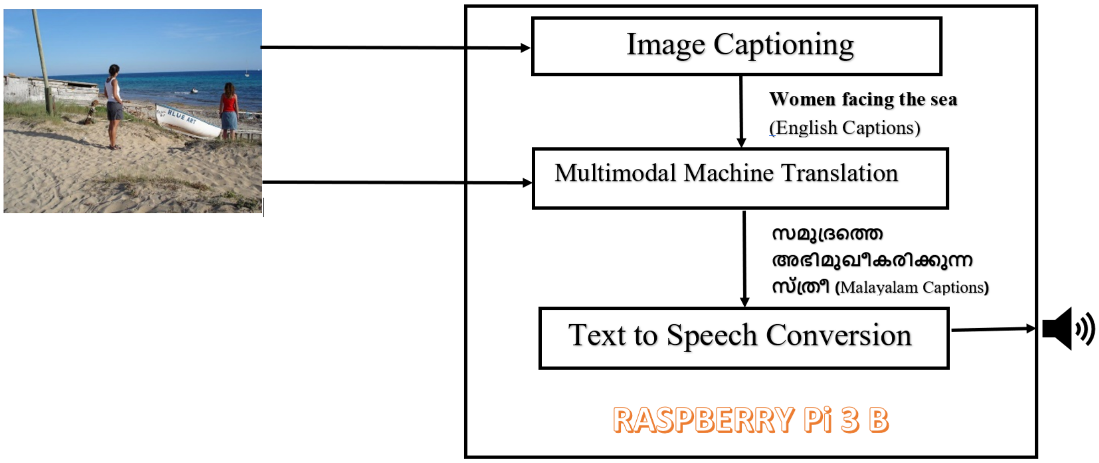

Figure 2. AI -Transl Design Approch

1. IoT sensor network
IoT components used in this project are Raspberry Pi 3 B, Pi cam, Microphone, and a button. Raspberry Pi 3 B acts as a microprocessor on this project. The camera connected with Pi can be triggered with a push-down button. Image captured using Pi cam is later processed to detect objects and to generate meaningful captions and their translation. This translation is converted into speech using Google cloud text to speech API and voice is outsourced through the headphone.The detailed system architecture is shown in figure 3
2. Image captioning
Image captioning models are constructed on top of sequential encoder-decoder circuits[17]. Image captioning models incorporate two sub-modules. An image model and the language model. These two models are combined using injecting architecture In injecting architecture, image encodingS and word embeddings are mixed and fed into an RNN network. A pre-trained convolutional neural network, like Visual Group Geometry 16 (VGG 16) or Inception network, is used to extract image features. These vectors are passed through a CNN Encoder network to create image encodings. Captions are preprocessed and tokenized using Natural Language Toolkit(NLTK). A new vocabulary is created out of it. Word embeddings are created using this vocabulary. The language model(decoder) is build using GRU layers. This model intakes image encodings and word embeddings as input and generate captions Bahdanau attention and teacher enforcing methods are utilized on this model for better results.
3. Multimodal machine translation
Multimodal machine translation[18] incorporates one or more contexts to produce effective translations. Multimodal machine

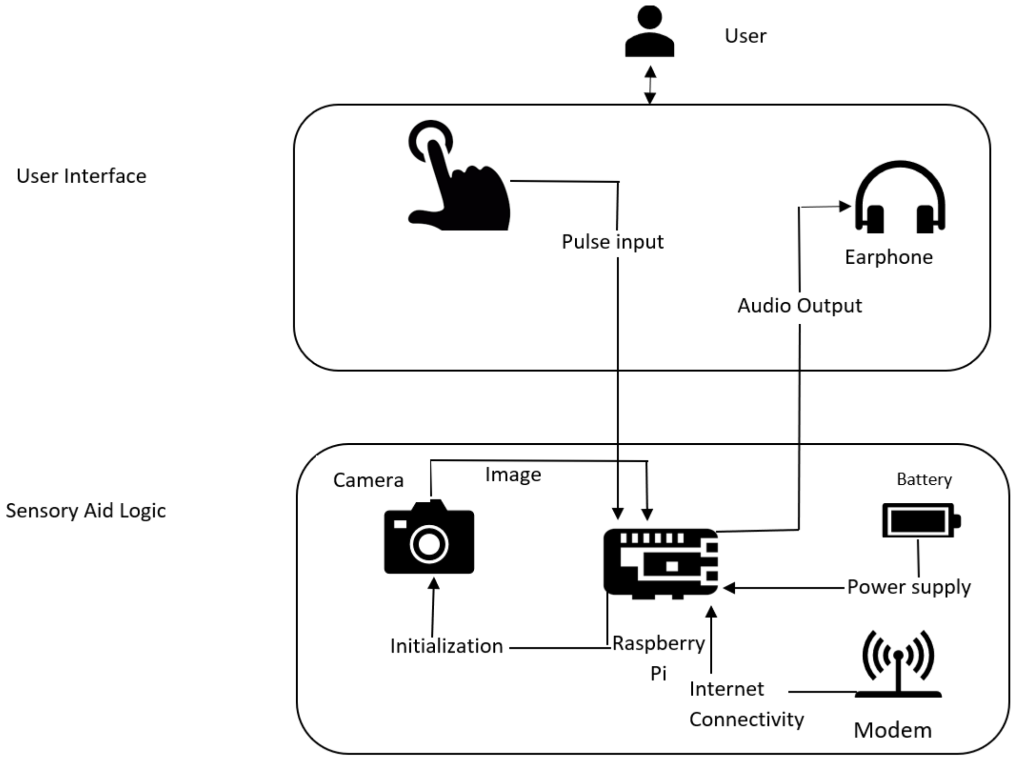

Figure 3. AI - Transl System Architecture

translation model architecture is very much similar to classical neural machine translation architecture. Both models operate on a sequential 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 pretrained VGG network is utilized to extract global and local features from the images. Image encodings generated from these features are utilized to initialize the decoder hidden state. Our model consolidates a Bidirectional RNN on the encoder part and doubly-attentive RNN on the decoder part[13]. As a result, this model applies separate attention to visual features and caption word embeddings.

Validation

This sensory aid makes use of two prominent methods in multimodal machine learning- multimodal image captioning and multimodal machine translation. So this project involves training and testing of two models using different datasets. For both tasks, we are using 2 different sets of datasets to evaluate our aid's performance.

1. Image captioning For image captioning, we have selected MS COCO (Microsoft Common Objects in Context) as our primary dataset. This dataset is commonly used for object detection, image segmentation, and image captioning. Coco dataset comprises 330 k images from 80 object categories and 93 stuff categories(objects without clear boundaries like the sky, street so on). Among these, 220 k of them are annotated. It also includes 5 captions per image. Different versions of the Coco dataset are available. This research uses the coco dataset released in 2014.
Flicker 30k dataset is used as the secondary dataset because of the gigantic size of the coco dataset. Flickr's 30 k dataset contains 31 k images collected from Flickr. This dataset consolidates 5 captions representing each image, which makes them suitable for image captioning at low resources.

2. Multimodal machine translation

In multimodal translation, we have chosen Malayalam Visual Genome (MVG) as our primary dataset. This dataset contains images, English captions, and corresponding Malayalam captions. This dataset comprises 29K images for training, 1K for development and 1.6K for testing. Images used in VG are taken from the MS COCO dataset. MVG dataset was released in 2021 as part WAT challenge.
Multi30 k dataset is used as the secondary dataset. It is used to convert English captions to European languages like German,

French. This dataset contains images of 30 k images taken from the flicker dataset, its English captions, and its German translations.

For the training and testing of these networks, we use tensorflow and pytorch with one NVIDIA TESLA P100 GPU

The most commonly used quantitative metrics for evaluating the performance of image captioning and multimodal machine translation tasks are: BLEU score(Bilingual Evaluation Understudy score) This score is used for comparing generated sentences to one or more reference sentences. This is used as a primary metrics to evaluate image caption generation and multimodal machine translation. Other than that RIBES score(Rank-based Intuitive Bilingual Evaluation Score) and METEOR scores (Metric for Evaluation of Translation with Explicit ORdering) are used to check the translation quality.

References

1. Aqel, M. O. A. *et al.* Review of recent research trends in assistive technologies for rehabilitation. In *2019 International Conference on Promising Electronic Technologies (ICPET)*, 16–21 (2019).

2. Ackland, P., Resnikoff, S. & Bourne, R. World blindness and visual impairment: despite many successes, the problem is growing. *Community Eye Health* **30**, 71 (2017).

3. Bourne, R. R. A. *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. *Lancet Global Health* **5**, e888–e897 (2017).

4. Senjam, S. S., Foster, A. & Bascaran, C. Barriers to using assistive technology among students with visual disability in schools for the blind in Delhi, India. *Disability and Rehabilitation: Assistive Technology* (2020).

5. Dabre, R., Chu, C. & Kunchukuttan, A. A survey of multilingual neural machine translation. *ACM Comput. Surv.* **53** (2020).

6. Liu, D., Ma, N., Yang, F. & Yang, X. A survey of low resource neural machine translation. In *2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 39–393 (2019).

7. Bhowmick, A. & Hazarika, S. An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. *Journal on Multimodal User Interfaces* **11**, 1–24 (2017).

8. Khan, M. A., Paul, P., Rashid, M., Hossain, M. & Ahad, M. A. R. An ai-based visual aid with integrated reading assistant for the completely blind. *IEEE Transactions on Human-Machine Systems* **50**, 507–517 (2020).

9. Amritkar, C. & Jabade, V. Image caption generation using deep learning technique. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–4 (2018).

10. Hengle, A., Kulkarni, A., Bavadekar, N., Kulkarni, N. & Udyawar, R. 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)*, 1109–1116 (2020).

11. Karmel, A., Sharma, A., Pandya, M. & Garg, D. IoT based Assistive Device for Deaf, Dumb and Blind People. *Procedia Comput. Sci.* **165**, 259–269 (2019).

12. Zhang, Z. *et al.* Neural machine translation with universal visual representation. In *International Conference on Learning Representations* (2020).

13. Calixto, I. & Liu, Q. Incorporating global visual features into attention-based neural machine translation. 992–1003 (2017).

14. Calixto, I., Liu, Q. & Campbell, N. 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)*, 1913–1924 (Association for Computational Linguistics, Vancouver, Canada, 2017).

15. Laskar, S., Singh, R., Pakray, D. P. & Bandyopadhyay, S. English to hindi multi-modal neural machine translation and hindi image captioning (2019).

16. Song, Y. *et al.* Enhancing neural machine translation with dual-side multimodal awareness. *IEEE Transactions on Multimedia* 1–1 (2021).

17. Hani, A., Tagougui, N. & Kherallah, M. Image caption generation using a deep architecture. In *2019 International Arab Conference on Information Technology (ACIT)*, 246–251 (2019).

18. Sulubacak, U. *et al.* Multimodal machine translation through visuals and speech. *Machine Translation* **34** (2020).