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

Lekshmy H O
amenp2ari20020@am.students
.amrita.edu

Dr.Swaminathan
swaminathanj@am.amrita.edu

Amrita School Of Engineering
Amrita Vishwa Vidyapeetham

September 2021

Outline

Introduction

Basics

- Assistive technology mainly focused on developing gadgets, frameworks and equipments.
- Significant piece of research activities aimed at developing assistive devices for the visually impaired.
- Over 70 percent of the blind population lives in multilingual developing countries.

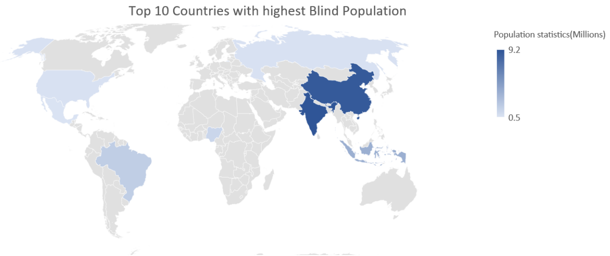

Figure: Top countries with blind population

Current Scenario

- A huge assortment of vision assistive gadgets are available.
- Vision aids operates on IOT sensory networks that works on computer vison and natural language processing principles.
- Major part of aids are customized for English speaking users.
- People in multi lingual developing countries often face hardship in using aids due to language barrier.
- Some aids incorporate Machine translations modules for customerization.

Common Implementation Techniques

- Vision aids commonly accommodates techniques like object detection and image captioning.
- Statistical machine translation and Neural machine translation systems are adopted for translation purpose.
- SMT is a "rule-based" MT method, it uses parallel bilingual text corpora for translation.
- NMT provides more accurate translation by accounting the context of each word.

Neural Machine Translation

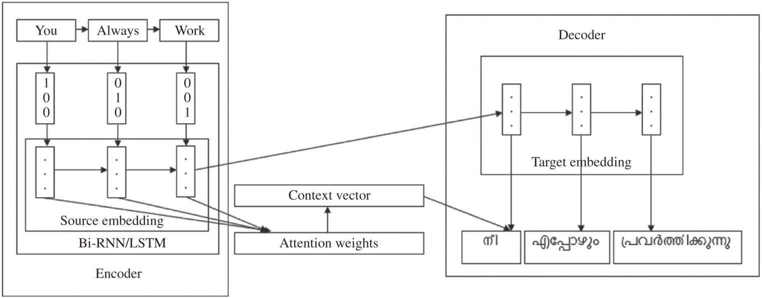

Figure: NMT - Architecture

Drawback of SMT and NMT based Sensory Aids

- SMT and NMT doesn't work well on low resource languages.
- Translations produced by SMT and NMT are literal, which eventually leads to misinterpretations.

Problems with Translation API

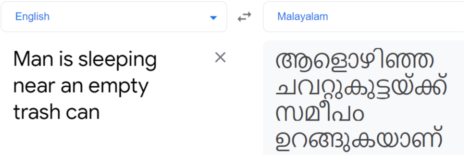

Figure: Google Translator

- Translation API suffers low translation quality for low resource languages.
- No security or confidentiality for data.
- High Pricing.

How to Overcome current issues?

- By incorporating Multi modal machine learning techniques in translation.
  - Multi-modal concept combines textual and visual features to improve the translation quality.
  - Multi modal machine translation works well on low resource language like Indian dialects.

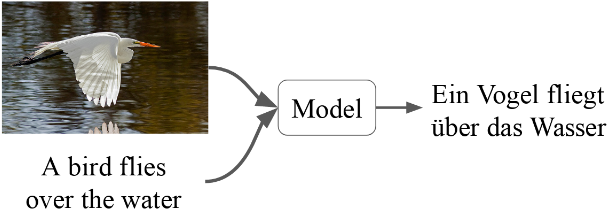

Figure: Multi modal machine translation on European dialects

Related works

Smart Cap: A Deep Learning and IoT Based Assistant for the Visually Impaired [3]

- Incorporated more advanced features like image captioning, face recognition and OCR for text identification.
- Operated using audio inputs.
- Drawback: Customized for English speaking users, Ignores foreign language speaker.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [9]

- Uses a convolutional neural network for feature extraction.
- LSTM is used to decode features into a sentence.
- Soft attention mechanism is incorporated to improve the quality of the caption
- Drawback:Attention adds more weight parameters to the model, which can increase training time .

Related work

Doubly-Attentive Decoder for Multi-modal Neural Machine Translation [1]

- Make use of image as additional modality for neural machine translation.
- Image features extracted using transfer learning is utilized to initialize the decoder.
- Drawback:Neglected semantic interactions between context vectors.

Multi-modal neural machine translation with deep semantic interactions [8]

- A bi-directional attention network for modeling text and image representations
- Co-attention network for refining text image context vectors.
- Drawback:Experiments conducted on English to French, English to Czech,Ignored low resource Asian languages.

Related work

M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training. [6]

- Model that combines multilingual pre-training and multimodal pre-training into a unified framework via multitask pre-training.
- Introduces Multimodal Code-switched Training.
- Drawback:Not a generalised model.

Improved English to Hindi Multimodal Neural Machine Translation [4]

- Make use of phrase pairs injection approach.
- SMT-based phrase pairs are augmented with the original parallel data to improve low-resource language pairs translation.

Related Work

Related Work Summary

- Vision based sensory device works on Computer Vision and natural language processing principles.
- Multi modal machine translation can generate efficient translations in low resource language.
- Majority of Multi modal machine translation researches are going on European languages.
- MMT are build on top of Encoder-decoder circuit with attention mechanism.
- Traditional MMT incorporates spatial visual features through a separate visual attention mechanism.

Research Gap

Research Gap

- An affordable effective assitive device for blind people in multilingual country is still lacking.
- Efficient Translation techniques for low resource Indian languages are still lacking.
- Translations on Dravidian languages is an understudied area due to lack of training data.

Solution

- Multimodal machine translation intakes more than modality for effective translation.
- Multimodal machine translation is the most efficient translation method for low resource Dravidian languages.
- Multimodal machine translation can be utilized in vision aids .

Proposed Method

In brief

- Develop a multi-linguistic sensory aid for blind individuals than can ,
  - Produce proficient textual descriptions from images.
  - Can translate descriptions into low-asset Indian dialects like Malayalam.
- Three different functionalities are integrated inside this device.
  1. IOT Sensor network
  2. Image Captioning
  3. Multi modal machine translation

IOT Sensor NeTWORK

- IoT components used in this project are Raspberry Pi 3 B, Pi cam, headphone, and a push down button.

Proposed Method

Proposed Model

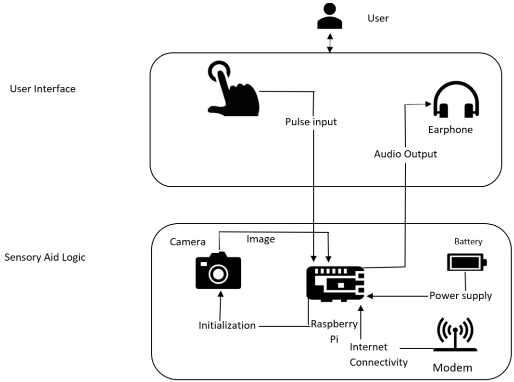

Figure: Proposed Model

- Image captured using Pi cam is processed to detect objects and to generate meaningful captions and their translation.

Proposed Method

WORK FLOW

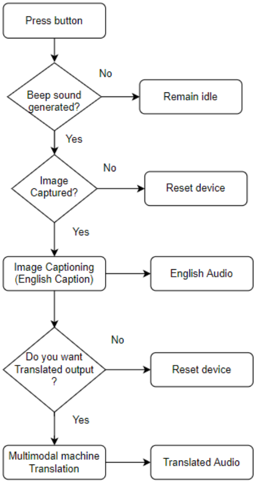

FIGURE: Conceptual workflow of proposed solution

Proposed Method

Image captioning

- Image captioning models are constructed on top of sequential encoder-decoder circuits with attention mechanism.
- Image features are extracted using transfer learning techniques.Encodings are generated from it.
- Language model intake image encodings and word embeddings to generate captions.

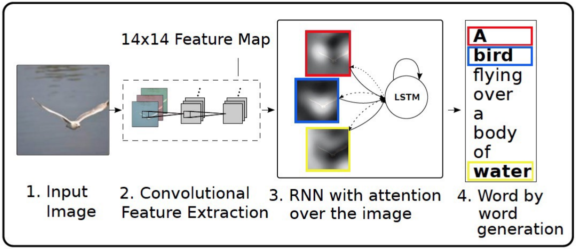

Figure: Image captioning Model

Proposed Method

Multimodal Machine Translation

- Multimodal machine translation incorporates one or more contexts to produce effective translations.
- Prime focus is given to
  - To study the impacts of inclusion of image as an additional context on neural machine translation.
  - To analyse the performance of MMTs on low resource Dravidian language translations.
  - To make comparative study on translation quality of MMT and NMT on low resource Dravidian language .
- Participation in WAT(Workshop on Asian Translation) 2022 Challenge

Proposed Method

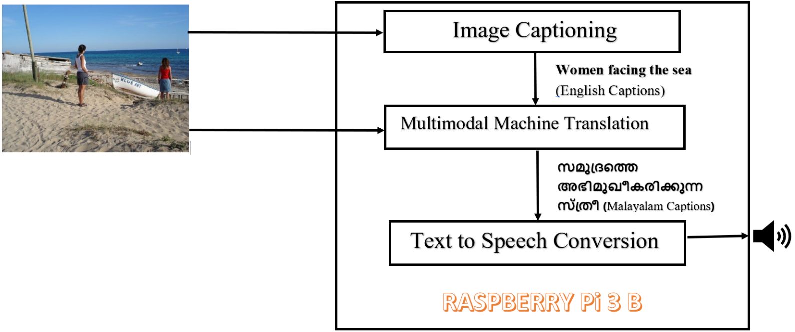

Figure: Work flow of MMT Model

- MMT model intakes image and English captions as input and translate captions into Malayalam.

Proposed Method

- Model is implemented using encoder-decoder circuit with LSTM cells.
- Image features are extracted using Transfer Learning
- Model uses a VGG -16 for feature extraction.
- Decoder hidden state is initialized using image encodings and embedding vectors.

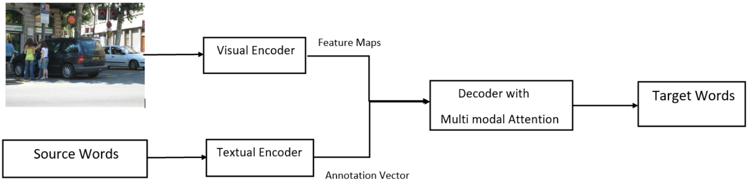

Figure: Multi Modal Machine Translation Model

Current Status

Current Status

- Model is trained for image captioning task.
- Tested Speech Recognition module.
- Completed Text To Speech Conversion.
- Performed manual error correction on VG Malayalam data set.
- Multi Modal Machine Translation with basic encoder decoder circuit completed.
- Writing paper titled 'English -Malayalam Bilingual assitive aid using Multimodal machine learning '

Validation

Platforms Used for Validation

- Final IoT project is validated in real time using Raspberry pi and connected devices.
- In training phase it is validated using following Frame works
  1. Tensor flow
  2. Tensor flow Lite
  3. Pytorch
- GPU used for validation is NVIDIA Tesla K80 GPU with Google Colab
- Main Packages
  1. PyAudio
  2. SpeechRecognition
  3. gTTS(Google Translate's Text-to-Speech API)
  4. Gensim
  5. Natural Language Toolkit

Dataset

Image Captioning

1. MS COCO (Microsoft Common Objects in Context)[5]
   - MS COCO dataset comprises 330 k images from 80 object categories and 93 stuff categories.
   - 220 k of images are annotated.
   - Dataset consolidates 5 captions representing each image.
   - Commonly used for object detection, image segmentation, and image captioning,pose estimation.
   - Size:25 GB

Dataset

Multimodal machine translation

1. Malayalam Visual Genome[7]
   - This dataset contains images, English captions, and corresponding Malayalam captions.
   - Dataset comprises 29K images for training, 1K for development and 1.6K for testing.
   - MVG dataset was released in 2021 as part WAT challenge.
2. Multi30k[2]
   - This dataset comprises 30 k images taken from the Flickr dataset, its English captions, and its German and French translations.

Validation

Performance Metrics

- Quantitative metrics for evaluating the performance of image captioning and multi modal machine translation tasks are
  1. BLEU score(Bilingual Evaluation Understudy score)
     - BLEU score is used for comparing generated sentences to one or more reference sentences.
  2. METEOR scores (Metric for Evaluation of Translation with Explicit ORdering
     - Used to check the translation quality.

Summary

- To develop a multilingual sensory aid for blinds.
- Multi modal machine translations can be utilised to convert the captions into low resource Indian language like Malayalam.
- Image is used as an extra modality to improve the performance of the model
- Double attentive encoder- decoder circuit produce relatively good image captioning and Multi modal machine translation systems.

Action Plan

Module	Subparts	Status	Expected date of Completion
Image captioning	Caption generation	Completed	
	Caption to Audio	Completed	
Multimodal machine Translation- Flicker	Feature Extraction	Completed	
	Developing Encoder – decoder circuit	Completed	
	Caption to Audio	Completed	
Writing First paper – Bilingual Assistive device for blinds with Multimodal machine translation	Image captioning	Completed	
	Multimodal machine Translation- VG Malayalam	Completed	
Multimodal machine Translation- Visual Genome Malayalam	Feature Extraction	Completed	
	Developing Encoder – decoder circuit	80%-Completed	10/10/2021
	Caption to Audio	Completed	
IoT Sensor Network	Integrating Sensor network		20/10/2021
	Deploying pretrained models		22/10/2021
	Testing models		25/10/2021
Writing second paper AI-Transl: A Multilingual Assistive Device			30/10/2021
Participation in WAT 2022 MMT Challenge			2/2/2022

Figure: Action plan

References I

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)*, pages 1913–1924, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia.
Multi30K: Multilingual English-German Image Descriptions.
arXiv, May 2016.

References II

Amey Hengle, Atharva Kulkarni, Nachiket Bavadekar, Niraj Kulkarni, and Rutuja Udyawar.
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)*, pages 1109–1116, 2020.

Sahinur Laskar, Rohit Singh, Dr. Partha Pakray, and Sivaji Bandyopadhyay.
Improved english to hindi multi-modal neural machine translation and hindi image captioning, Jul 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Zitnick.
Microsoft coco: Common objects in context.
05 2014.

References III

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan.
M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training.
ArXiv, 06 2020.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash.
Hindi visual genome: A dataset for multi-modal english to hindi machine translation.
Computación y Sistemas, 23(4), 2019.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai.
Multi-modal neural machine translation with deep semantic interactions.
Inform. Sci., 554:47–60, Apr 2021.

References IV

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio.
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.
arXiv, Feb 2015.