AI-TransI: Multilingual Assistive Device Using Multi Modal Machine Learning

Lekshmy H O[1] and Dr. Swaminathan J[2,*]

[1] Dept of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri
e-mail: amenp2ari20020@am.students.amrita.edu
[2] e-mail: swaminathanj@am.amrita.edu

Submitted on July 31, 2021; accepted xxx xx, 2021

ABSTRACT

Context. As per statistics, around 36 million people in the world are blind and more than 253 million people suffer from visual impairment. For decades scientists are trying to find aids that can support visually impaired people in day-to-day activities. These aids include navigational aids, educational aids, and so on. A couple of assistive devices are available in the market. Most of these devices are built on top of Computer Vision and Natural Language Processing. The majority of these devices are developed by multinational companies rooted in developed countries, as a result, most of these aids are developed for English language users. Customizing an aid for a particular linguistic group is costlier, in a multilingual country like India. As a result, 90 percent of blind people living in developing countries don't use navigational aids other than walking sticks. To increase affordability, some assistive devices make use of translation APIs and modules to translate English into the regional language, but these translations are not up to the mark when comparing with the original text. This paper leverages multimodal machine learning techniques to create an efficient and effective sensory aid for the visually impaired. Multimodal image captioning helps to produce more accurate captions, while multimodal machine translations help to translate the generated captions to another language by making use of visual features extracted from images. This aid can outperform regular assistive aids and can produce more meaningful captions and their translations in the regional language.
Key References: (Grover et al. 2021), (Khan et al. 2020), (Parida et al. 2019), (Hengle et al. 2020), (Sulubacak et al. 2020), (Ma et al. 2019),(Rajan et al. 2009),(Zhang et al. 2020b),(Laskar et al. 2019),(Singh et al. 2021)
Aims. To create a multilingual assistive aid for blind and visually impaired people using IoT technology and multimodal machine learning techniques like Multimodal Image captioning and Multimodal Machine translation.
Key References: (Khan et al. 2020), (Nishihara et al. 2020), (Wang & Hasegawa-Johnson 2020), (Zhang et al. 2020a),(Anto & Nisha 2016),(C & Ganesh 2015),(Sunil et al. 2011)
Methods.

This project focus on creating an assistive device for blind people and visually impaired people with the help of IoT components and machine learning algorithms. The proposed system's primary task is to describe surroundings or scenes to blind people in the regional language with the help of visual features. For that, it makes use of computer vision and natural language processing algorithms. This model incorporates hardware and software components. The hardware part includes a Raspberry pi, Wi-Fi- connected Camera, and a headset. Image captured using the camera is processed inside the Raspberry pi using deep learning models. These models are trained for multimodal caption generation and multimodal machine translation. Caption generation models understand the content of images and convert them into an embedding vector. This act as a feed to a language model for generating captions Transfer learning methods are using to extract features from the images. The language model works on the encoder-decoder circuits with an attention mechanism. Output from the encoder circuit and feature extractor is combined in the decoder part to generate English captions. These captions and images become input for the multimodal translation model. This model translates captions to regional Indian languages like Malayalam or Hindi. Multimodal machine translation makes use of a sequential encoder-decoder circuit with an attention mechanism. Global features extracted using transfer learning methods are used to initialize the hidden state of the encoder. The decoder will combine visual features and text corpus to generate translated captions in the regional language. These translations are later converted into speech using text to speech API.
Key References: (Song et al. 2021),(Amritkar & Jabade 2018), (Yao & Wan 2020),(Tan et al. 2020),(Rojan et al. 2020),(Sunil et al. 2011),(Liu et al. 2021),(Baltrušaitis et al. 2019)
Validation. (Optional):We are making use of the BLEU(Bilingual Evaluation Understudy) score as a metric to validate the quality of translation, this metric is commonly used for comparing a generated sentence with an reference sentence.
Key References: (Meetei et al. 2019), (Laskar et al. 2020), (Laskar et al. 2020),(Delbrouck & Dupont 2017)

Key words. Assistive device – Multi modal machine translation –Multi modal Image Captioning – Blind assistive device—visually impaired assistive device

References

Amritkar, C. & Jabade, V. 2018, in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1–4

Anto, A. & Nisha, K. K. 2016, in 2016 International Conference on Emerging Technological Trends (ICETT), 1–6
Baltrušaitis, T., Ahuja, C., & Morency, L.-P. 2019, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41, 423
C, A. V. & Ganesh, A. 2015, in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1565–1570
Delbrouck, J.-B. & Dupont, S. 2017, 910–919
Grover, M., Rathi, R., Manchanda, C., Garg, K., & Beniwal, R. 2021, in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 354–359
Hengle, A., Kulkarni, A., Bavadekar, N., Kulkarni, N., & Udyawar, R. 2020, in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 1109–1116
Khan, M. A., Paul, P., Rashid, M., Hossain, M., & Ahad, M. A. R. 2020, IEEE Transactions on Human-Machine Systems, 50, 507
Laskar, S., Khilji, A., Pakray, D. P., & Bandyopadhyay, S. 2020
Laskar, S., Singh, R., Pakray, D. P., & Bandyopadhyay, S. 2019, English to Hindi Multi-modal Neural Machine Translation and Hindi Image Captioning
Liu, X., Zhao, J., Sun, S., Liu, H., & Yang, H. 2021, Information Fusion, 69, 73
Ma, J., Qin, S., Su, L., Li, X., & Xiao, L. 2019, in 2019 International Conference on Asian Language Processing (IALP), 199–204
Meetei, L., Singh, T. D., & Bandyopadhyay, S. 2019, 181–188
Nishihara, T., Tamura, A., Ninomiya, T., Omote, Y., & Nakayama, H. 2020, 4304–4314
Parida, S., Bojar, O., & Motlicek, P. 2019, 175–180
Rajan, R., Sivan, R., Ravindran, R., & Soman, K. 2009, in 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, 439–441
Rojan, A., Alias, E., Rajan, G. M., Mathew, J., & Sudarsan, D. 2020, in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 161–165
Singh, Y. P., Ahmed, S. A. L. E., Singh, P., Kumar, N., & Diwakar, M. 2021, Journal of Physics: Conference Series, 1854, 012048
Song, Y., Chen, S., Jin, Q., et al. 2021, IEEE Transactions on Multimedia, 1
Sulubacak, U., Çağlayan, O., Grönroos, S.-A., et al. 2020, Machine Translation, 34
Sunil, R., Manohar, N., Jayan, V., & Sulochana, K. G. 2011, in 2011 Annual IEEE India Conference, 1–6
Tan, L., Li, L., Han, Y., et al. 2020, in 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), 63–69
Wang, L. & Hasegawa-Johnson, M. 2020, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 1560
Yao, S. & Wan, X. 2020, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online: Association for Computational Linguistics), 4346–4350
Zhang, C., Yang, Z., He, X., & Deng, L. 2020a, IEEE Journal of Selected Topics in Signal Processing, 14, 478
Zhang, Z., Yu, H., Zhao, H., Wang, R., & Utiyama, M. 2020b, Accurate Word Representations with Universal Visual Guidance

*Dissertion Supervisor