Using Deep Learning Techniques for Image-to-Bangla Natural Language Text Generation to Improve Mobility for the Visually Impaired in Bangladesh

Ashiqur Rahman Saron1, b), Hasin Almas Sifat1, c), Md. Mortuza Ahmmed2, d), Md Saef Ullah Miah3, e) and Goh Kah Ong Michael4, a)

1Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh

2Department of Mathematics, American International University-Bangladesh, Dhaka, Bangladesh

3Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh

4Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450, Bukit Beruang, Melaka, Malaysia

*a) Corresponding author: michael.goh@mmu.edu.my*

*b)22-48697-3@student.aiub.edu  
c) 22-48679-3@student.aiub.edu*

*d) mortuza@aiub.edu*

*e) saef@aiub.edu*

**Abstract.** In Bangladesh, nearly 950,000 people are blind, and 27 million face vision loss. Visually impaired individuals encounter significant challenges in navigation and situational awareness. Additionally, Bangladesh ranks 63rd in English proficiency globally, highlighting the necessity of solutions based on the local language to effectively reach the general population. This project aims to develop a model that converts real-time images into Bangla natural language text and then transforms it into speech, enhancing mobility and independence. By using a custom dataset to train the model, we aim to improve accuracy in Bangla natural language generation, particularly benefiting those who are not proficient in English. Deep learning models will be implemented for better image-to-Bangla text generation, ensuring higher accuracy. The system captures the surrounding environment, processes images using deep learning-based object detection, and translates the detected surroundings into Bangla text. This text is then converted into speech using text-to-speech translation techniques, providing auditory feedback to the user. The proposed system enhances accessibility by offering real-time contextual information about the environment, enabling safer and more efficient movement in daily life. It is designed to assist individuals with both partial and full blindness, improving maneuverability and guidance.

# InTRODUCTION

Navigating urban and rural environments without vision presents substantial difficulties, from avoiding obstacles to interpreting contextual cues [1]. The usage of artificial intelligence (AI) has shown potential in addressing these challenges, but many existing solutions are primarily developed for translation of text [2]. Bangladesh ranks 61 globally out of 116 countries in English proficiency and this indicates the importance of native-language-based inventions [3]. Recent advancements in deep learning have enabled significant improvements in image captioning and scene understanding [4,5]. While natural language generation in low-resource languages like Bangla is gradually gaining attention but challenges remain in resource collection [6]. Integrating these technologies with speech synthesis systems can provide real-time audio feedback, facilitating independent mobility for the visually impaired [7]. In recent years, computer vision techniques have been widely applied to develop assistive systems for the visually impaired and it includes wearable cameras, smart glasses, and mobile applications [8]. However, most of these systems either lack support for native languages or do not offer contextual verbal guidance which makes them less effective for non-English speakers. Moreover, research shows that incorporating localized languages in assistive technologies not only improves user experience but also significantly enhances adoption among rural and underserved populations [9]. By leveraging custom datasets and Bangla language models, our system aims to address these gaps by tailoring the technology to fit the cultural and linguistic context of Bangladesh. Furthermore, with the growing popularity of transformer-based architectures like Vision Transformers [10] and multilingual text generation models [11], there is an increasing opportunity to build robust image captioning systems for underrepresented languages. By focusing on real-time processing and high-accuracy Bangla captioning, this research aspires to improve environmental awareness, mobility, and overall quality of life for people with partial or complete blindness in Bangladesh.

# RELATED WORKS

The problem of aiding visually impaired individuals in Bangladesh, has garnered significant attention in recent years, especially with the proliferation of assistive technologies such as Braille-based solutions and machine learning applications. One of the early works in this domain has explored mBRAILLE, an Android application designed to aid visually impaired students in learning Braille. While it proved useful for tactile learning, it lacked the capacity for real-time object recognition and environmental description, which are crucial for navigating real-world environments [12]. Advancements in machine learning have led to systems designed for real-time Braille and sign language detection. Tanveer et al. utilized CNNs to detect and convert visual data into Braille and sign language. However, their system did not include description generation, which limits its ability to provide holistic environmental awareness for users [13]. Shakib et al. presented a machine learning-based system using YOLOv7 for real-time face detection and distance measurement, but their focus on object detection alone left the broader issue of environmental scenario detection unexplored [14]. IoT-based solutions have also been explored, such as Ghosh, who employed TensorFlow object detection with Raspberry Pi and Coral USB Accelerator. While effective for object detection, the system lacked description generation and audio feedback, which are critical for providing real-time contextual information to the user [15]. Similarly, Rajatha developed a smart cap using machine learning to assist visually impaired individuals, but it lacked both descriptive capabilities and support for Bengali as a communication medium [16].

Further developments in wearable devices, such as those proposed by Hasan et al. focused on real-time object detection using YOLOv3 for navigation assistance. While it provided valuable navigation aids, it did not include description generation or auditory feedback [17]. Islam et al. proposed a specialized dataset for enhancing object recognition, yet the integration of such datasets into practical applications remains a challenge [18]. Deep learning-based solutions using pre-trained models like SSDLite and MobileNetV2 showed promise but were hindered by their reliance on English for communication and the absence of conversational audio descriptions [19]. Oion et al. combined object detection with NLP to generate audio descriptions, but this system remained limited to detecting and describing objects rather than providing a full environmental context [20].

Hasan et al. focused on creating systems for specific scenarios, such as assisting visually impaired visitors at the Liberation War Museum in Bangladesh. However, these systems were restricted to a specific set of objects and did not support the Bengali language, making them less useful for the wider population [21].

Recent studies, such as those by Hossain et al. explored footpath detection for visually impaired individuals, but like many other systems, it focused on detection rather than comprehensive environmental descriptions [22]. Furthermore, while image captioning systems for low-resource languages, such as Kazakh, have been explored, Bengali remains underutilized. Arystanbekov et al. highlighted the potential of image captioning for the visually impaired in Kazakh, but the Bengali language remains underexplored in this area [7]. Lastly, Masud et al. explored the use of visual attention mechanisms for generating captions in Bengali, yet its integration with audio descriptions for the visually impaired remains an unexplored area, highlighting a significant gap in assistive technology research [6].

# MeTHODOLOGY

## Data Preparation

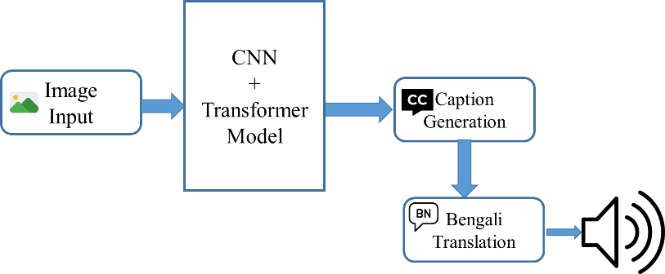
To train the image captioning model effectively, a preprocessing pipeline was applied to the Flickr8k caption data (Figure 1). Each line in the caption file was parsed to extract image names and captions, with trailing identifiers (e.g., #0) removed for consistency. Captions with fewer than 5 tokens or exceeding a set length were discarded to ensure uniform input. Start and end tokens were added to valid captions to define sequence boundaries. Images with no remaining valid captions were excluded. The final image-caption pairs were shuffled and split into 80/20 training and validation sets, ensuring clean, consistent, and properly formatted data for vocabulary building and sequence modeling.

|  |  |
| --- | --- |
| **FIGURE 1.** Data preparation | **FIGURE 2.** Model training and caption generation |

## Model Training & Caption Generation

The proposed image captioning system shown in Figure 2 integrates a convolutional neural network (CNN) with a transformer-based encoder-decoder framework. Images are preprocessed (decoded, resized, normalized) and passed through a frozen EfficientNetB0 model (excluding the top layer). The output feature maps are reshaped into a 2D sequence format suitable for transformer input. A custom encoder block processes the CNN features using a feedforward Dense layer followed by Multi-Head Self Attention. Residual connections and Layer Normalization ensure training stability. Captions are vectorized using a TextVectorization layer and embedded using a custom PositionalEmbedding layer, which includes both token and positional encodings scaled by embedding dimension. The decoder uses causal self-attention to process input captions and cross-attention to align with encoder outputs. Feedforward layers and dropout are used to refine the decoded representation, followed by a Dense output layer predicting vocabulary tokens. A custom Image Captioning Model integrates the CNN, encoder, and decoder. During training, each image is paired with multiple captions. Loss and accuracy are calculated across all caption variants using a masked sparse categorical cross-entropy loss. The model is trained using the Adam optimizer with a custom warm-up learning rate scheduler and early stopping.

## Bangla Text Generation

A sample of 216 images representing various aspects of Bangladesh was collected for the purpose of testing the performance of the caption generation model [24]. The images were not pre-labeled to evaluate the model's ability to generate accurate captions independently. Each generated caption was then assigned a score based on its relevance to the Bengali description of the image. The captions were subsequently converted from text to speech, forming the foundation of our proposed model. This model aims to assist the visually impaired population by providing audio descriptions of their surroundings. Figure 3 showcases the architecture of the speech generation.

**FIGURE 3.** Image to speech generation

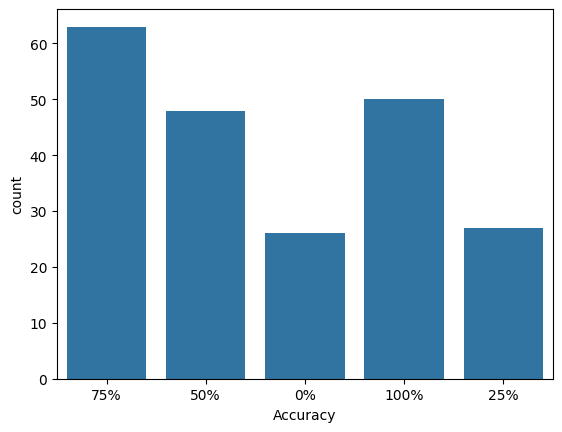
# RESULTS

The given test images were given to the model and Bengali caption was generated and based on their accuracy a score was given. Some sample output image Bengali caption is shown in Table 1.

**Table 1.** Bangladeshi images and output caption

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Input Image** |  |  |  |  |  |
| **Bengali Caption** | গোলাপি শার্ট পরা একটি মেয়ে এবং মেয়ে হাসছে | একদল লোক একদল লোকের সারিতে বসে আছে | টুপি আর সানগ্লাস পরা একটা ছোট বাচ্চা হাসছে | একটি নীল জ্যাকেট লাল ভবনের সামনে দাঁড়িয়ে আছে | একজন পুরুষ এবং একজন মহিলা গিটার বাজাচ্ছেন |
| **Caption Accuracy** | Mostly Accurate (100%) | Almost Accurate (75%) | Half Accurate (50%) | Slightly Accurate (25%) | Not Accurate (0%) |

From the test images, over 60 images were found to have an accuracy of 75%. More than 45% of the images showed 50% accuracy. Nearly 50 images achieved 100% accuracy in caption generation. Around 25 images were found to give inaccurate results, with accuracy rates of 0% and 25%. The results are shown in Figure 4.



**Figure 4.** Accuracy of the generated captions

**Table 2** presents the most frequently used words in the generated captions of Bangladeshi images. The table highlights not only the frequency but also the diversity of vocabulary used by the model. This analysis provides insight into the model’s linguistic tendencies and the dominant visual themes it captures from Bangladeshi contexts. Frequent usage of words like “man,” “woman,” “shirt,” and common prepositions indicates the model’s focus on people and attire, which are often central to the images. Understanding this distribution helps evaluate the descriptiveness and cultural relevance of the generated captions.

**Table 2.** Most frequent words

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| একটি | মধ্যে | এবং | হচ্ছে | পুরুষ | শার্ট | আছেন | উপর, | এর | মহিলা |
| 508 | 132 | 123 | 112 | 94 | 88 | 75 | 74 | 56 | 43 |

# Discussion

The approach of utilizing CNN and Transformer for caption generation and then followed by translation into Bengali using the Google Translate API and after that converting the text into speech with Google Text-to-Speech (TTS) represents an innovative solution for aiding visually impaired individuals. Transformer models for image captioning are highly effective in generating context-aware captions that describe the visual surroundings. By using these models, we can ensure that the captions generated are accurate, concise, and capable of providing users with relevant details about their environment. Once we have generated the captions, the translation into Bengali using the Google Translate API is crucial for making the system accessible to the target population. Google Translate offers a robust solution for translating between English and Bengali and it ensures that the system can cater to the needs of native Bengali speakers, which is vital for inclusivity. Finally, the use of Google Text-to-Speech (TTS) allows for seamless auditory feedback, enabling visually impaired users to understand the surrounding environment without relying on visual cues. The TTS system can convert the translated text into natural-sounding speech and further enhances the practicality of the system by providing real-time, accessible information.

Overall, this approach provides a promising solution that combines state-of-the-art NLP, image captioning, and speech synthesis technologies, making it a promising method to enhance the mobility and independence of visually impaired individuals.

# Limitation

Despite the promising performance of the proposed image captioning model, several limitations remain. The transformer-based decoder and multi-head attention mechanisms introduce high memory and computation requirements, especially during training, limiting scalability to very large datasets or deployment on low-resource devices. Although multiple captions per image are used during training, the model may still generate generic or repetitive captions and lacks the diversity seen in human-generated descriptions. The model for feature extraction restricts the model’s ability to adapt image representations to the captioning task, potentially limiting performance. The fixed vocabulary learned from the training corpus may cause the model to struggle with out-of-vocabulary words, uncommon objects, or new contexts not present in the training set. The model lacks the ability to incorporate real-world knowledge or context beyond the image and caption pair and it can be crucial for accurately describing complex or abstract scenes.

# Conclusion

In conclusion, the proposed system represents a significant advancement in assistive technology for the visually impaired in Bangladesh. By combining transformers for accurate image caption generation, Google Translate for seamless Bengali translation, and Google Text-to-Speech for real-time speech output, the system provides a comprehensive solution for enhancing the situational awareness of blind and low-vision individuals. The system not only addresses the need for a local-language solution but also leverages powerful AI tools to create an intuitive and user-friendly experience. This combination of technologies offers a promising direction for future research and development in accessibility tools for the visually impaired.

# References

1. R. Manduchi and S. Kurniawan, "Mobility-related accidents experienced by people with visual impairment," Insight: Research and Practice in Visual Impairment and Blindness, vol. 4, 2011.M. P. Brown and K. Austin, Appl. Phys. Letters **85**, 2503–2504 (2004).
2. M. H. Haque, M. F. Hossain, and A. F. Hossain, "Machine and web translator for English to Bangla using natural language processes," Daffodil International University Journal of Science and Technology, vol. 5, no. 1, pp. 53–61, 2010.
3. EF Education First, "EF English Proficiency Index: Bangladesh," EF EPI, [Online]. Available: <https://www.ef.com/wwen/epi/regions/asia/bangladesh/>. [Accessed: Apr. 5, 2025]
4. K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in Proc. Int. Conf. Machine Learning (ICML), PMLR, 2015.
5. P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 6077–6086, doi: 10.1109/CVPR.2018.00636.
6. A. Masud, M. B. Hosen, M. Habibullah, M. Anannya, and M. S. Kaiser, "Image captioning in Bengali language using visual attention," PLOS ONE, vol. 20, 2025, doi: 10.1371/journal.pone.0309364.
7. B. Arystanbekov, A. Kuzdeuov, S. Nurgaliyev, and A. Varol, "Image captioning for the visually impaired and blind: A recipe for low-resource languages," in Proc. 2023 IEEE EMBC, 2023, pp. 1–4, doi: 10.1109/EMBC40787.2023.10340575.
8. K. G. Shanthi, "Smart vision using machine learning for blind," Int. J. Adv. Sci. Technol., vol. 29, pp. 12458–12463, 2020.
9. C. Mnyanyi, "ICT-based assistive technology for empowering persons with visual impairment," J. Issues Pract. Educ., vol. 14, 2023, doi: 10.61538/jipe.v14i2.1214.
10. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
11. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meet. Assoc. Comput. Linguist., Online, 2020, pp. 8440–8451.
12. L. Nahar, et al., "Design of a Braille learning application for visually impaired students in Bangladesh," Assistive Technology, vol. 27, no. 3, pp. 172-182, 2015.
13. T. M. Mahin, J. F. Islam Zerin, and N. T. Ava, "Real Time Braille & Sign Language Detection Using Artificial Neural Networks," M.S. thesis, Department of Electrical and Electronics Engineering (EEE), Islamic University of Technology (IUT), Board Bazar, Gazipur-1704, Bangladesh, 2024.
14. M. S. Shakib, M. E. Ali, R. Khisa, Kamaruzzaman and S. Bhowmik, "Machine Learning Based Smart System for Blind and Visually Impaired People," 2024 2nd International Conference on Information and Communication Technology (ICICT), Dhaka, Bangladesh, 2024, pp. 204-208, doi: 10.1109/ICICT64387.2024.10839742.
15. A. Ghosh, "Assistive Technology for Visually Impaired using Tensor Flow Object Detection in Raspberry Pi and Coral USB Accelerator," 2020.
16. R. Rajatha, "Smart Cap For Visually Impaired People Using Machine Learning," Journal of Science, Computing and Engineering Research, vol. 7, no. 8, 2024.
17. M. Z. Hasan, S. Sikder, and M. A. Rahaman, "Real-time computer vision based autonomous navigation system for assisting visually impaired people using machine learning," 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), IEEE, 2022.
18. M. T. Islam, et al., "A Dataset for Crucial Object Recognition in Blind and Low-Vision Individuals' Navigation," arXiv preprint arXiv:2407.16777, 2024.
19. R. B. Islam, et al., "Deep learning based object detection and surrounding environment description for visually impaired people," Heliyon, vol. 9, no. 6, pp. e16809, 2023.
20. M. S. Oion, et al., "A Machine Learning Based Image to Object Detection and Audio Description for Blind People," 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, 2023.
21. M. M. Hasan, et al., "Automatic Detection and Recognition of Object to Help Visually Impaired People while Visiting Liberation War Museum in Bangladesh," 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, 2021.
22. M. I. Hossain, J. Anjom, and R. I. Chowdhury, "Towards walkable footpath detection for the visually impaired on Bangladeshi roads with smartphones using deep edge intelligence," Array, vol. 100388, 2025.
23. K. Anitha Kumari, C. Mouneeshwari, R. B. Udhaya, and R. Jasmitha, "Automated image captioning for Flickr8K dataset," in Proceedings of the International Conference on Artificial Intelligence, Smart Grid and Smart City Applications (AISGSC 2019), Cham: Springer, 2020, pp. 639–648.
24. M. A. Jishan, K. R. Mahmudand A. K. A. Azad, “Bangla Natural Language Image to Text (BNLIT)”. Zenodo, Aug. 21, 2019. doi: 10.5281/zenodo.3372752.