Personalized English Amharic Medical Image Caption and Speech Generation for Visually Impaired Patients Using Vision Transformer Fused with LLM

Dawit Shibabaw[1,*], Vukosi Marivate[2], Munir Awol[3], and Tesfa Tegegne[4,*]

[1]Bahir Dar Institute of Technology, Bahir Dar University, Faculty of Computing, Bahir Dar, Ethiopia

[2]University of Pretoria, Computer Science, Pretoria, South Africa

[3]Tikur Anbessa Hospital, Oncology, Addis Ababa, Ethiopia

[4]Bahir Dar Institute of Technology, Bahir Dar University, Faculty of Computing, Bahir Dar, Ethiopia

*corresponding authors email: BDU1500754@bdu.edu.et and Tesfa.Tegegne@bdu.edu.et

Abstract

Access to medical information is critical for healthcare equity, particularly for visually impaired citizens and low-resource language speakers. Our goal is to create a model that enables visually impaired individuals to access their medical image results, by convert the text into an audio message, and translate generated captions into local languages to understand their medical results in their mother tongue. Developing algorithms that can generate captions, translating into Amharic and generate speech from images is a major goal of our study by fussing computer vision and Generative AI. In this study, following the design science approach, the data were gathered from the Tikur Anbessa specialized hospital, Addis Ababa University, and the data annotation was carried out by a domain expert. We preprocessed the data to make suitable for models. The work presents a novel approach model fusion such as Vision Transformer (ViT)-GPT2, VIT-Llama2, and VGG16-LSTM architectures for medical image captioning. The model is designed to generate detailed captions for radiologists, translate the generated caption into Amharic, and speech for visually impaired patients. Among the models' ViT-Llama2 model generate high-quality caption and robust feature extraction, ensures precise, context-aware captions. Experiments demonstrate the effectiveness of this method, VIT-Llama2 achieving a high BLEU score of 0.633 in image captioning and enhanced usability and accessibility. The system is deployed as a user-friendly application that accepts medical images as input, processes them through the models, outputs textual captions, translates generated caption into Amharic, and speech. This model bridges the gap in medical accessibility for low-resource language speakers, empowering visually impaired individuals and understand their medical image results.

Keyword: Medical image, Caption, visually impaired, Translation to Amharic, VIT, speech Generation, Low Resource language.

1. Introduction

Getting a system to automatically display the content of an image and a natural language description/report is one of the most challenging issues in computer vision. This paper uses model fusion approaches to generating image captions, such as vision transformer and large language models(LLM)(Preetham & Krishnan, 2024). Automatic image captioning is developed predominantly for high-resourced languages compared to low-resource languages(Cho & Oh, 2023). Therefore, generating captions for medical images for low resource language are crucial for clinicians and patients(Barreto et al., 2023), (Razafinirina et al., 2024). Thus, the study attempts to develop an automatic image

captioning system for medical images and translate the caption into Amharic language. In addition, generating an image caption (text) is translated or converted to speech for poorly educated (semi-educated) and visually impaired patients/people(Cho & Oh, 2023). In this study, developing algorithms that can generate captions, translating into local languages, and generating speech from generated captions is a major goal of our study by combining computer vision and Generative AI (Krishnakumar et al., 2020). Access to medical image results is essential for patients with blind or visually impaired and low-resource language speakers in their mother tongue (Chen et al., 2023).

2. Method

2.1. Dataset

This study used secondary data, from the Tikur Anbessa Hospital, Addis Ababa University, Ethiopia. The image and textual data repositories were in different databases. so, first step, we extracted textual data having both demographic and clinical information of the chronic cancer patients. This textual data contains the unique medical record number (MRN) of each patient. In this data, we have the patient medical record number (MRN) of each patient used to extract images from the image repository. Most of the time, a patient is prescribed to have a radiology test if the stages of the cancer are stage 3, or stage 4 or the severity level of the tumor is malignant. We have used a total dataset of 3430 images. Each image has 2-10 corresponding captions. In this study x-ray, MRI, and CT scan images are included for patients diagnosed with liver, lung, breast, colorecatal, and cervical cancers.

2.2. Mapping text and image data

In this dataset, a single image contains 2–10 sentences. For example, if MRN is 100522 we name it specific image as 100522.jpg, then for a list of sentences/captions for this image 100522.jpg#1, 100522.jpg#2, 100522.jpg#3 etc. until captions ends as shown in Fig 1. if not remove captions (or images) that do not have corresponding images (or captions). as shown in Fig 1 take 100522.jpg image_id and take a list of caption's caption_id such as 0, 1, 2, 3, 4 and caption.

Table 1: mapping caption with image.

Image ID	Caption ID	Caption
100522.jpg	0	startseq There is right large heterogeneously enhancing upper lobe mass...
100522.jpg	1	startseq There is right superior sulcus mass leading to bony destruction...
100522.jpg	2	startseq There is middle lobe bronchiectasis and nodularity...
100522.jpg	3	startseq Conclusion Spinal canal narrowing as a result of bony invasion...
100522.jpg	4	startseq Conclusion Known cervical ca with multiple metastases...

2.3. Data Augmentation (Training Only)

To enhance generalization and robustness, the following augmentations were applied during training:

Random horizontal flips (same count as of the original, but augmented on-the-fly)

2.4. Model Fusion Strategy

In this study, we integrate a language model with a visual model to improve the generation of captions for medical images(Zhu et al., 2024). In particular, a feature level fusion approach was employed. The initial step was employing a Vision Transformer (ViT) to extract visual elements from the medical image. Textual features (such sentence embeddings) were simultaneously acquired from a language model that had already been trained, such as all-MiniLM-

V2. Concatenating these two features sets one from the text and one from the image created a single representation. To create the final caption, this fused representation was subsequently passed into a decoder model (such as Llama-2, GPT-2, and LSTM). With this method, the model may simultaneously use language context and visual content to provide captions that are more accurate and clinically appropriate.

3. Experimental Result and discussion

We have employed model fusion such as Generative pretrained transformer-2(GPT-2) fussed with Vision transformer (Vit), Llama2 fused with ViT, and VGG16 Fussed with Long short term (LSTM). We evaluated the fusion model by consensus-based image description evaluation (CIDEr), bilingual evaluation understudy (BLEU), To evaluate the translation. We also used Metric Evaluation of Translation with Explicit ORdering (METEOR), and recall-oriented understudy for gisting evaluation (ROUGE). Table 1 shows an overall result of the model:

Table 1: The overall result of the experiment

Fused models	BLEU	ROUGE-2	ROUGE-L	ROUGE-1	METEOR	CIDEr
VIT-GPT-2	0.543	0.550	0.245	0.254	0.524	0.424
VIT-Llama-2	0.633	0.350	0.445	0.154	0.324	0.236
VGG16-LSTM	0.342	0.320	0.130	0.163	0.264	0.425

The model is designed to generate detailed captions for radiologists, translate generated captions into local languages to access their medical result by their mother tongue, and speech for visually impaired patients.

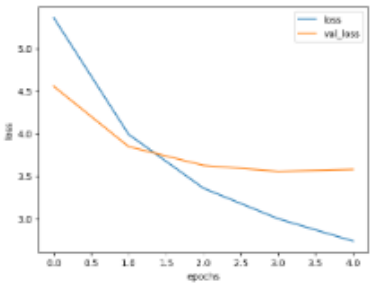


Fig 4: training and validation loss.

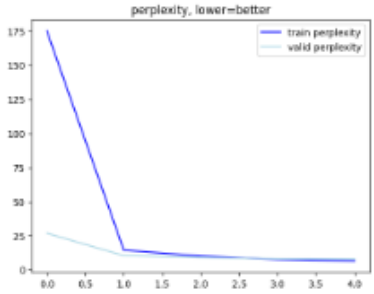


Fig 5: perplexity, lower=better.

EarlyStopping This stops training when the validation loss stops improving, to prevent overfitting and underfitting.

L2 regularization, also known as weight decay or ridge regularization, is a method that penalizes big weights in the loss function in order to reduce model complexity and avoid overfitting.

Perplexity: Measures how well a model predicts a sequence of text. It is formally defined as the exponential of the average negative log-likelihood of a sequence. Mathematically:

$$\text{Perplexity} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)}$$

Where:

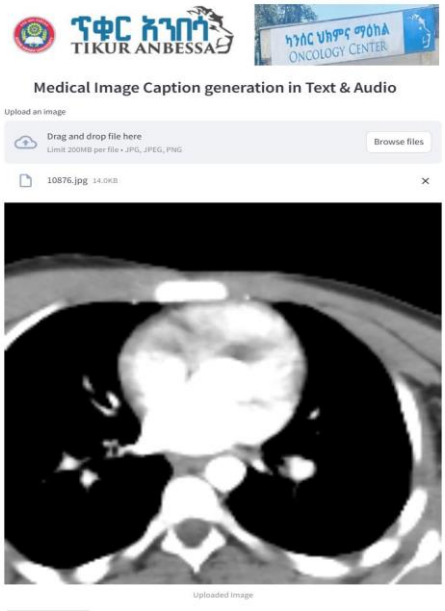
- N: The total number of words (or tokens) in the sequence.
- $P(w_i)$: The probability of the i-th word in the sequence, as predicted by the model.

3.1. Implementation Details

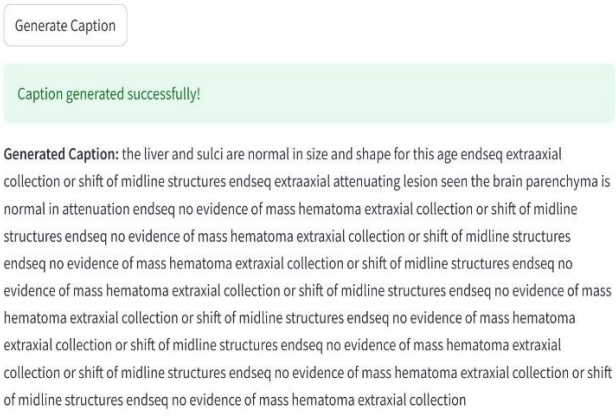
We saved our fused model, Vision Transformer (ViT) fussed with LLaMa-2, as my_model.keras to facilitate real-world deployment. Additionally, Python's pickle module was used to serialize the corresponding captions and store them as a file called caption.pkl.

Figure 1: Implication details

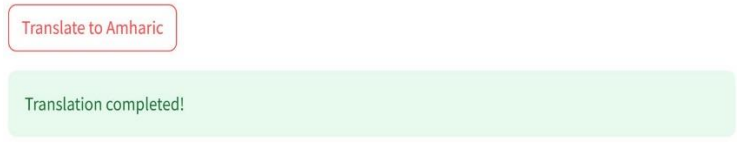
Demonstration: Step1-Step4 Outputs



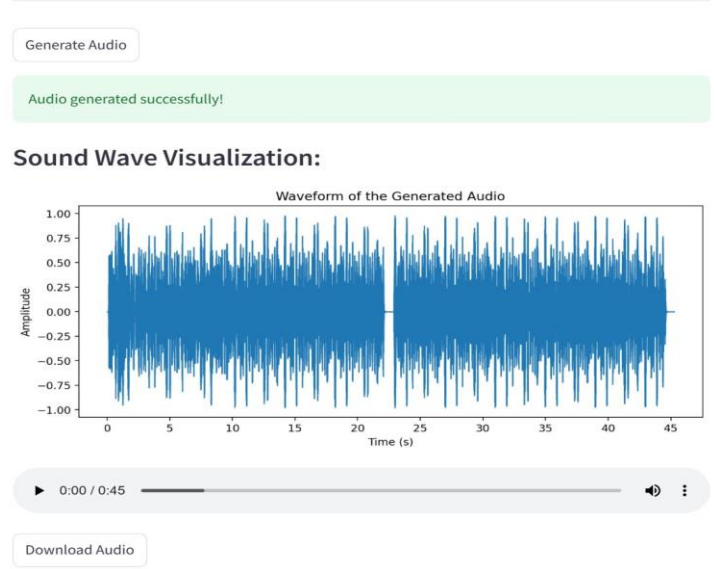
Step 1: GUI accept Image



Step2: generated caption (English)



Step 3: Generated caption translated Into Amharic



Step 4 generated Speech_visualized wave

Figure 2: compression of model caption with actual caption

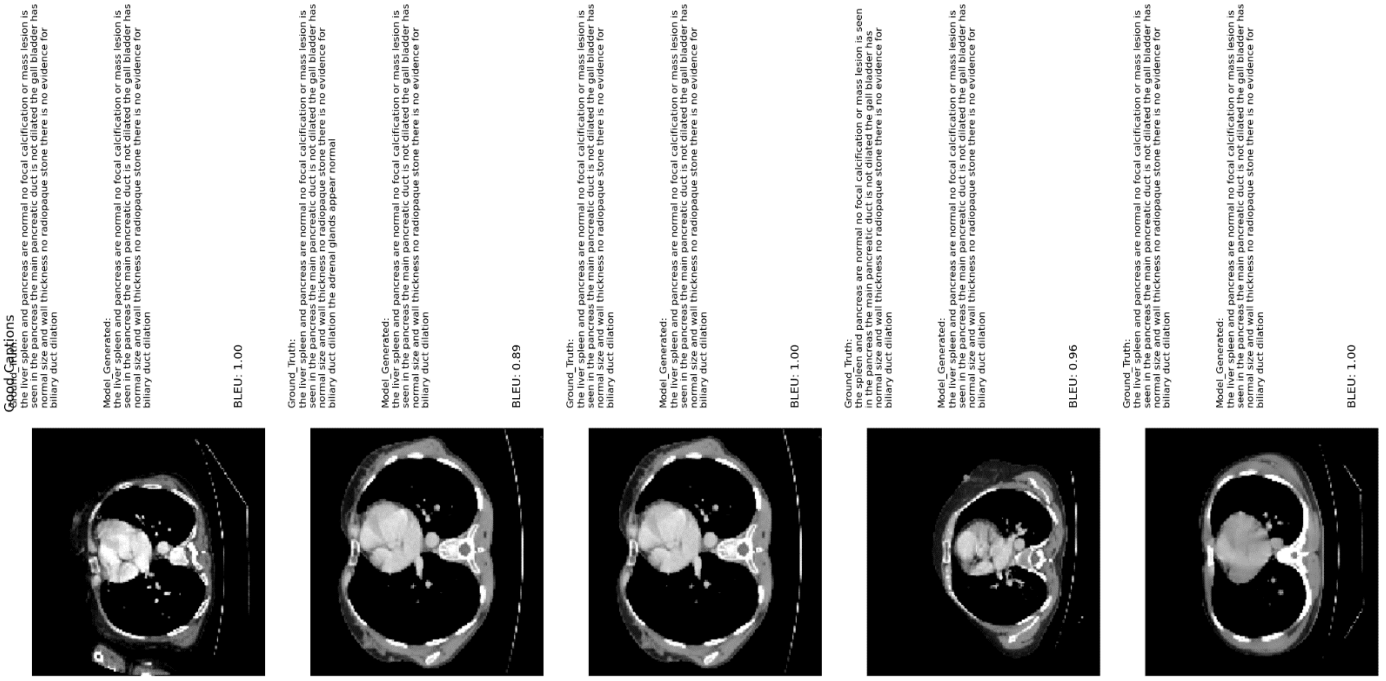
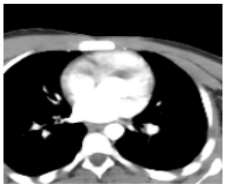


Table 2: Comparison ChatGPT-4/Gemini 2.5 pro Report vs Actual Report

Image	ChatGPT-4 Generated/Gemini 2.5 Report	Actual Report
	**Findings:** • The heart is visible in cross-section, and appears enlarged relative to the thoracic cavity. This may suggest cardiomegaly, although proper size measurements are needed. • Lung fields are visualized. Both left, and right lungs appear to show preserved aeration; no clear signs of mass, consolidation, or effusion are observed in this single slice. • Aorta and great vessels are partially visualized; no clear aneurysmal dilation or dissection signs noted in this frame. • Aorta and great vessels partially visualized; no dilation noted. • Ribs and thoracic bones show no gross fracture. **Impression:** • No acute abnormality on this axial CT chest slice. • Prominent heart — recommend size correlation and clinical history. • Recommend full CT review for complete evaluation.	**Findings:** • 100522.jpg#0 : The right and left vertebral arteries arising from the bilateral subclavian arteries have small caliber, they pass through the transverse foramina and end up to communicate with the tortious vessels at C1-2 level. the intracranial segment of vertebralarteries is not visualized. The basilar artery is mainly a continuation of a branch of the left ICA. • 100522.jpg#1: Left maxillary sinus mucous retnetion cysts. • 100522.jpg#2: The glottis, supra- and subglottis spaces are unremarkable. No abnormal area of contrast enhancement. • 100522.jpg#3: Conclusion: Large, lobulated and infiltrative heterogenous left posterolateral deep neck space soft tissue mass with promeint vascular channels and internal phleboliths ; likely suggesting Hemangioma/ soft tissue venous malformation **Impression:** • 100522.jpg#4: Conclusion: Bilaterally small cervical vertebral arteries with basial artery supplied by left ICA

3.2. Clinically validation

As part of the clinical validation shown in Table 5, four radiologists from Tikur Anbessa Hospital evaluated the accuracy of the generated medical image captions. Each radiologist rated the captions based on their clinical correctness and diagnostic relevance using a 5-point scale: 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent.

Table 3. Fused model Generated Caption accuracy evaluation by four clinical radiologists at Tikur Anbessa Hospital.

Metric	Radiologist 1	Radiologist 2	Radiologist 3	Radiologist 4
Caption Accuracy	3 (Very Good)	4 (Excellent)	3 (Good)	3 (Very Good)

- Mean Caption Accuracy Score:(3 + 4 + 3 + 3)/4 = 3.25

The average rating of 3.25 from clinical radiologists demonstrates that the generated captions are generally very good. These results indicate that the model-generated captions are largely clinically reliable.

3.3. Translation pipeline into Amharic

In this study, the generated English medical image captions are translated into Amharic and researchers used a Transformer-based Machine Translation (TMT)(Gezmu, 2018), (Asefa & Assabie, 2025) framework which incorporates a multilingual encoder-decoder architecture and self-attention. The model is pre-trained on large-scale multilingual corpora and fine-tuned for English-Amharic translation, allowing for efficient transfer learning of this low-resource language. While delivers fluent and generally accurate outputs for broad text, reliability falls for domain-specific content due to inadequate Amharic training data(Hadgu et al., 2020).

4. Conclusion

This research demonstrates the potential of model fusion ViT-GPT2, ViT-Llama2, and VGG16-LSTM architectures for generating medical image captions for radiologist, translate generated captions into local language to access their medical image results by their mother tongue, and generate speech for visually impaired patients. The model ensures accurate captioning for radiologists and intuitive speech for visually impaired patients, and translating into low resource language to promote inclusivity in healthcare. Deployment of this technology will offer a transformative solution for medical accessibility, addressing the needs of under-served populations. We compared result of proposed model with CNN-RNN and Generative AI, while the proposed model (ViT, Llama 2, chatgpt2, LSTM, and VGG16) works better than others in medical image captioning.

5. Author contributions statement

T.T. and V.M. supervised the study, M.A., G.T., and B.T. annotated the Data with his team, D.S. preprocessed the data, both image and captions, developed the model, and wrote the report. Study concept design: T.T. and V.M. All authors are involved in the following activities: Manuscript revision, data access and verification, final review and approval of manuscript before submission.

6. Data and source code Availability Statement:

Due to patient privacy and the rules of the centers' institutional review boards, the datasets created and analyzed during this investigation are not publicly available; however, they can be obtained from the corresponding author upon reasonable request. You can find the codes/demo to visit the GitHub repository: you can find the codes/projects in the GitHub repository: Click here.

Reference

Asefa, S. H., & Assabie, Y. (2025). Transformer-Based Amharic-to-English Machine Translation With Character Embedding and Combined Regularization Techniques. *IEEE Access*, *13*(December 2024), 1090–1105. https://doi.org/10.1109/ACCESS.2024.3521985

Barreto, A. G., de Oliveira, J. M., Gois, F. N. B., Cortez, P. C., & de Albuquerque, V. H. C. (2023). A New Generative Model for Textual Descriptions of Medical Images Using Transformers Enhanced with Convolutional Neural Networks. *Bioengineering*, *10*(9). https://doi.org/10.3390/bioengineering10091098

Chen, Y., Lin, Y., Xu, X., Ding, J., Li, C., Zeng, Y., Xie, W., & Huang, J. (2023). Multi-domain medical image translation generation for lung image classification based on generative adversarial networks. *Computer Methods and Programs in Biomedicine*, *229*, 107200. https://doi.org/10.1016/J.CMPB.2022.107200

Cho, S., & Oh, H. (2023). Generalized Image Captioning for Multilingual Support. *Applied Sciences (Switzerland)*, *13*(4), 1–15. https://doi.org/10.3390/app13042446

Gezmu, A. M. (2018). *Neural Machine Translation for Amharic-English Translation*. https://doi.org/10.5220/0010383905260532

Hadgu, A. T., Beaudoin, A., & Aregawi, A. (2020). *EVALUATING AMHARIC MACHINE TRANSLATION*. 1–4.

Krishnakumar, B., Kousalya, K., Gokul, S., Karthikeyan, R., & Kaviyarasu, D. (2020). Image caption generator using deep learning. *International Journal of Advanced Science and Technology*, *29*(3 Special Issue), 975–980. https://doi.org/10.55041/ijsrem31987

Preetham, M., & Krishnan, M. (2024). A Multi-Modal Image Understanding and Audio Description System for the Visually Impaired People. *10th International Conference on Advanced Computing and Communication Systems, ICACCS 2024*, 1014–1019. https://doi.org/10.1109/ICACCS60874.2024.10716992

Razafinirina, M. A., Dimbisoa, W. G., & Mahatody, T. (2024). *Pedagogical Alignment of Large Language Models (LLM) for Personalized Learning : A Survey , Trends and Challenges*. 448–480. https://doi.org/10.4236/jilsa.2024.164023

Zhu, D., Chen, X., & Gan, Y. (2024). *A Multi-Model Output Fusion Strategy Based on Various Machine Learning Techniques for Product Price Prediction ARTICLE A Multi-Model Output Fusion Strategy Based on Various Machine Learning Techniques for Product Price Prediction*. *December*. https://doi.org/10.30564/jeis.v4i1.7566