VisionAssist : Visual Question Answering for Visually Impaired

*Anjana S,*Arya J,*Sandra V Jayaraj,*Sreelakshmi T Raj,**Dr. Sruthy Manmadhan

*UG Scholar,**Assistant Professor, Department of Computer Science and Engineering , N.S.S College of Engineering, Palakkad

PROBLEM STATEMENT

- Existing Visual Question Answering (VQA) systems are not designed to cater visually impaired users, lacking features for accessibility and usability.
- Traditional VQA techniques heavily depend on visual inputs, excluding users with impaired vision from interactive experiences.
- The lack of interpretability and transparency in VQA models hinders users' ability to comprehend the reasoning behind the system's answers.
- Current solutions for visually impaired individuals often rely on manual interventions or non-interactive tools, lacking real-time, context-aware responses to visual queries.

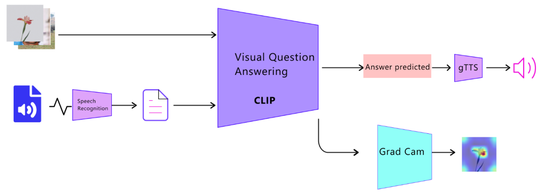
OBJECTIVES

- Develop a Visual Question Answering (VQA) system tailored to the needs of visually impaired individuals.
- Implement CLIP (Contrastive Language-Image Pre-training) and Vision Transformer (ViT), to enable efficient processing of visual and textual inputs for accurate responses.
- Integrate explainability features using techniques like Grad-CAM to enhance transparency and interpretability in the VQA system, providing users with insights into how answers are generated.

DATASET

- The VizWiz dataset contains 20,510 image-question pairs for training, 4,319 pairs for validation, and 8,000 pairs for testing.
- Blind individuals captured images and asked questions related to various daily-life scenarios.
- Multiple crowd workers answered these questions, resulting in 10 different answers per question to capture diverse perspectives.

METHODOLOGY



.

Fig 1 : Architecture of Proposed System

- Preprocess the dataset containing image-question-answer triplets.
- Utilized CLIP as the backbone for processing image and text inputs.
- Integrate Vision Transformer (ViT) variant into the CLIP architecture for image encoding.
- Implement fusion techniques within the CLIP-ViT architecture for effective multimodal feature combination.
- Obtain the final prediction.

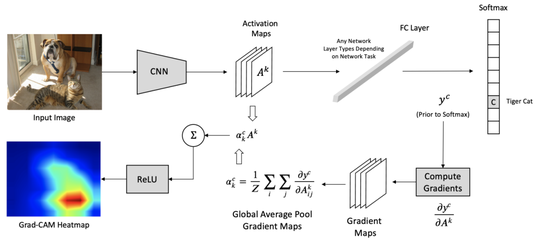


Fig 2 : Grad-CAM Architecture

- Converts logits into probabilities using softmax for decision-making.
- Computes gradients to assess the contribution of tokens.
- Derives attention maps via Grad-CAM for both image and text inputs.
- Generates heatmaps to visually represent critical features in images and questions.

RESULT AND ANALYSIS

- The proposed model achieved a test accuracy of 73.3%

Table 1 : Performance Analysis

Training Accuracy	76.21%
Validation Accuracy	75.52%
Test Accuracy	73.3%
Training Answerability Score	.81
Validation Answerability Score	.80
Test Answerability Score	.80

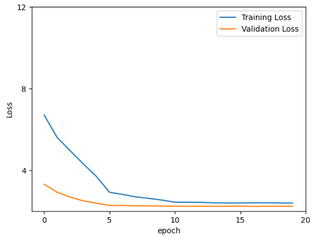


Fig 3 : Training Loss and Validation Loss graph

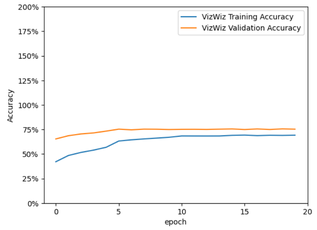


Fig 4 : Accuracy graph

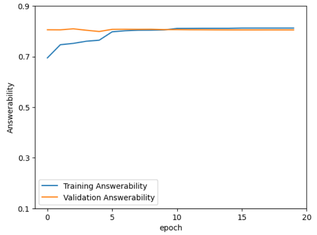


Fig 5 : Answerability Score graph

TOOLS

- PyTorch
- NumPy
- Matplotlib
- gTTS
- Captum

CONCLUSION

- Adapted VQA technology specifically for visually impaired users
- Applied interpretability techniques like Grad-CAM to provide transparency in decision-making, building user trust.
- Demonstrated robust fusion of textual and visual inputs using advanced techniques within the CLIP-ViT architecture, leading to improved response accuracy.

REFERENCES

[1] Deuser, Fabian, et al. ”Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model.” arXiv preprint arXiv:2206.05281 (2022).

[2] Chefer, Hila, Shir Gur, and Lior Wolf. ”Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[3] Selvaraju, Ramprasaath R., et al. ”Grad-CAM: Visual explanations from deep networks via gradient-based localization.” Proceedings of the IEEE international conference on computer vision. 2017.

[4] Tung, L., H. T. Nguyen, and M. L. Nguyen. ”Multi visual and textual embedding on visual question answering for blind people.” Neurocomputing 465 (2021): 451-464.

[5] Le, Tung, et al. ”Bi-direction co-attention network on visual question answering for blind people.” Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 12084. 2022.

[6] Mohith, S. Shiv, et al. ”Visual world to an audible experience: visual assistance for the blind and visually impaired.” 2020 IEEE 17th India Council International Conference (INDICON). IEEE, 2020.