Contextual Visual Reasoning Approach for Visually Impaired People

Rahad Arman Nabid

December 20, 2023

Abstract

Individuals with visual impairments or limited vision often face challenges in interpreting images, particularly those found on the web. This is especially true for complex images that require advanced reasoning for understanding. Despite these individuals constituting about 8% of the population, there is still a notable lack of tools providing contextual explanations of visual data. Current solutions, such as screen magnifiers, high-contrast software, screen readers, voice recognition software, adjustable screen brightness, color adjustment software, Braille displays, text-to-speech software, AI-powered image recognition tools, and optical character recognition software, are available to assist visually impaired people. However, while AI-powered image recognition tools give a general overview, they lack the capability to provide reasoning-based contextual explanations. Our project seeks to address this shortfall by creating a vision-language model that generates context-specific questions and relevant interpretations of visualization. We utilized a Visual Question Generation (VQG) model with a BLEU-1 score of 0.758092 and a BLEU-2 score of 0.579911, alongside a Visual Question Answering (VQA) model with 48.05% accuracy in generating answers. By combining these models, we developed a web application to gather feedback from visually impaired users. In our study involving 8 participants, all reported that the system significantly aided them in better understanding images.

1 Introduction

Individuals with visual impairments are those who experience either partial or complete vision loss that cannot be corrected to normal levels using glasses, contact lenses, or surgery. This term encompasses a wide range of visual experiences, from mild difficulties in viewing web images to more severe vision loss. A variety of diseases and conditions can lead to visual impairments, including Cataracts, Age-related macular degeneration (AMD), Glaucoma, Diabetic retinopathy, Refractive errors, and Retinal detachment [1]. The 2020-2022 National Eye Disease Prevalence Study (NEDPS) estimated that 24.1 million adults, or 8.3% of the population aged 40 and above, suffered from conditions such as refractive errors, cataracts, AMD, diabetic retinopathy, or glaucoma [2]. Consequently, the way visually impaired individuals perceive various visualizations is significantly different from those with normal vision, often struggling to interpret different contextual meanings from images. To aid in this, visually impaired people utilize various tools like screen magnifiers, high-contrast software, screen readers, voice recognition software, adjustable screen brightness controls, color adjustment software, Braille displays, text-to-speech (TTS) software, AI-powered image recognition tools, and optical character recognition (OCR) software [3]. Additionally, emerging technologies like AI-powered descriptions and sensory substitution devices are providing new possibilities for assistance [4].

In their paper, "A Smart Personal AI Assistant for Visually Impaired People," Felix et al. developed an Android application that is capable of object recognition in the user's surroundings and can perform text analysis on hard copy documents [5]. Similarly, Enamul et al. explored the development of a machine learning model in their study that is designed to generate visual explanations and summaries derived from charts, aiding in the interpretation of complex visual data [6]. In another significant contribution, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Kelvin et al. introduced a machine learning model that focuses on generating image captions. This technology is particularly useful for visually impaired individuals, as it provides them with descriptive narratives of images [7]. A study described a method for generating inclusive and dense descriptions from images, particularly useful in the context of webinars for visually impaired people. Daniel et al. describe methods involves detecting people in images, generating dense captioning, filtering these captions, and then using external classifiers for additional information like age and emotion. The process culminates in generating coherent summaries, providing a comprehensive context of the image for visually impaired users.

The majority of AI-powered tools offer a general overview to visually impaired individuals, but they lack the capability for reasoning-based explanations to enhance contextual understanding. Our project addresses this gap by developing a system that integrates a Visual Question Generation (VQG) model with a Visual Question Answering (VQA) model. This innovative system generates a series of contextually relevant questions by analyzing various features within an image. Additionally, the Visual Question Answer Generation model is employed to produce diverse types of reasoned answers for these questions. The primary objective of our project is not solely to build machine learning models but to answer two pivotal questions: 1) How can we design a contextual visual reasoning system by integrating VQG and VQA models? 2) How effective is this contextual visual reasoning approach in aiding visually impaired individuals with reasoning tasks?

2 Motivation:

Visually impaired individuals often face significant challenges in interpreting images on the web, leading to a sense of exclusion from this digital space. While existing tools like screen magnifiers, high-contrast software, screen readers, and AI-powered image recognition tools offer some assistance, they primarily provide a general overview without delving into the deeper, contextual understanding of content. This gap highlights the need for innovative solutions that go beyond basic recognition, offering reasoning-based explanations to foster a more inclusive and comprehensive web experience for visually impaired users. Our approach aims to address this need by introducing novel methodologies that enhance contextual comprehension, thereby bridging the accessibility gap and promoting a sense of belonging on the web for visually impaired users.

3 Background:

Visual impairments are not exclusive to a single condition but are symptoms that arise from various underlying health issues. Individuals with visual impairments may encounter a diverse array of difficulties, with the specific nature and extent of their challenges varying based on both the root cause of their vision loss and its severity. The provided figure illustrates typical examples of how people with visual impairments might perceive images, highlighting the broad spectrum of vision alterations experienced by this population. The figure-1 shows how normal people see vs visually impairment people see in real life.

Glaucoma: Glaucoma is typically characterized by a loss of peripheral vision, often described as "tunnel vision," where the central vision remains intact but the outer edges are blurred or missing.

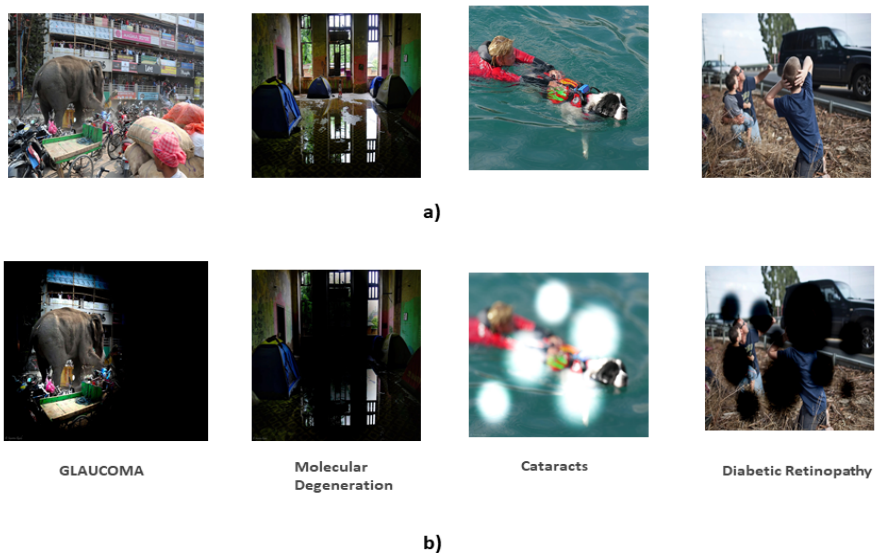

Figure 1: Perspectives of Perception: a) Clarity in Sight - The World Through Typical Vision b) A Different Spectrum - The World As Seen by Individuals with Visual Impairments

It is caused by damage to the optic nerve, usually from increased pressure in the eye. Glaucoma can occur at any age but is more common in older adults.

Macular Degeneration: Often mistakenly referred to as "Molecular Degeneration," Macular Degeneration affects the macula, the part of the retina responsible for clear vision in your direct line of sight. This can make it difficult to see fine details, whether you're looking at something close or far. The central vision becomes blurred while peripheral vision remains unaffected. Age-related macular degeneration (AMD) usually affects older adults and is a leading cause of vision loss in those aged 50 and older.

Cataracts: Cataracts cause clouding of the eye's lens, leading to a decrease in vision. It often develops slowly and can affect one or both eyes. Symptoms include faded colors, blurry vision, halos around light, trouble with bright lights, and trouble seeing at night. While it can occur at any age, it is most commonly associated with aging and is very common in older adults.

Diabetic Retinopathy: This disease is a result of prolonged high blood sugar levels causing damage to the blood vessels of the retina. It can lead to spots or dark strings floating in one's vision (floaters), blurred vision, fluctuating vision, dark or empty areas in your vision, and vision loss. Diabetic retinopathy often affects people who have had diabetes for a long time, typically adults.

Color Blindness: Color Blindness, or Color Vision Deficiency, is a condition that affects an individual's ability to see colors under normal lighting conditions. The most common form is red-green color blindness, where individuals find it challenging to distinguish between red and green hues. Symptoms can range from mild difficulty in differentiating colors to a more significant impairment where colors cannot be perceived at all. This condition is usually inherited and affects a significant proportion of the population, with males being more commonly affected due to its X-linked genetic pattern.

4 Methodology:

Individuals with visual impairments experience the world visually in ways that differ from those with typical vision. Challenges such as blurred images in cases of cataracts, or the appearance of black

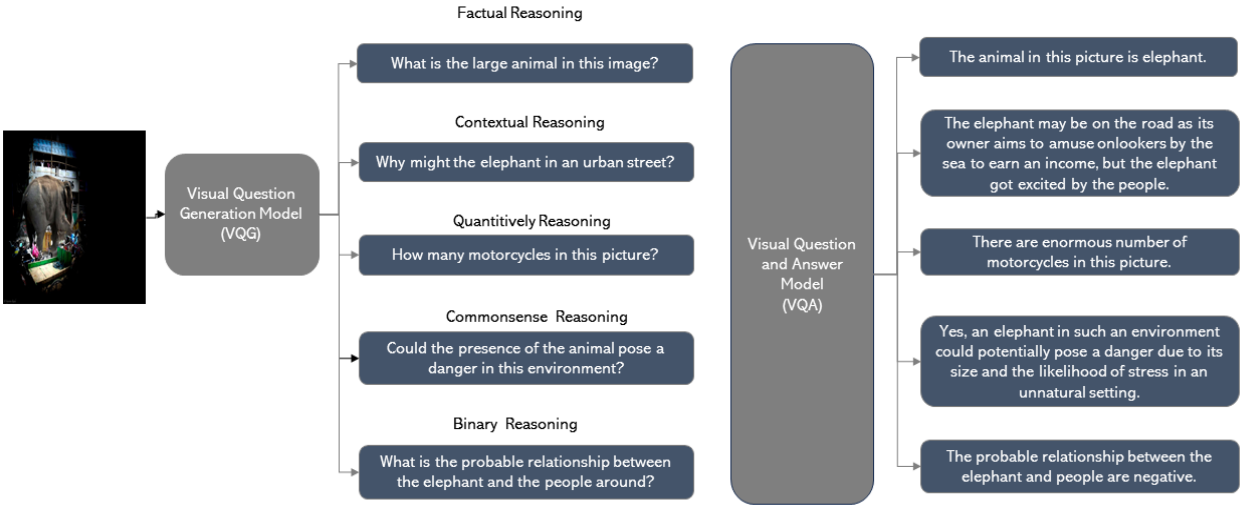

Figure 2: Contextual Visual Reasoning System

and white spots as seen in diabetic retinopathy and glaucoma, may cause them to miss critical details and make it difficult to discern and relate to various objects within an image. Standard systems offer features like chatbots for asking questions, but for those with visual challenges, these are not always sufficient due to the difficulty in visual exploration.

In our project, we propose that providing example questions may offer a starting point for visually impaired users to engage with the system more actively. To draw attention to the key elements in visualizations, we pose various reasoning questions, encompassing arithmetic, inductive, binary, common sense, and relational reasoning. This approach encourages visually impaired users to delve deeper and inquire further. Users are empowered to inquire based on provided examples and to pose their own reasoning questions, to which the system responds.

Our architecture incorporates two models: the Visual Question Generation (VQG) model, which generates initial questions, and the Visual Question and Answer (VQA) model, which responds to relevant queries. The overall system architecture is detailed in the figure- below.

4.1 Technical Details of Visual Question Generation Model:

4.1.1 Dataset for VQG:

The data preparation workflow illustrated involves enriching the Flickr8k dataset, which consists of 8,000 images, each with five descriptive captions, by transforming these descriptions into reasoning questions. This transformation is executed on a subset of 1,000 images, with a generative model creating five unique reasoning questions for each image, thus enhancing the dataset from mere descriptions to a form that necessitates cognitive reasoning to answer. This enriched dataset is likely intended to train a more advanced Visual Question Answering (VQA) model that can not only understand and describe visual content but also infer and reason about the context and elements within the images.

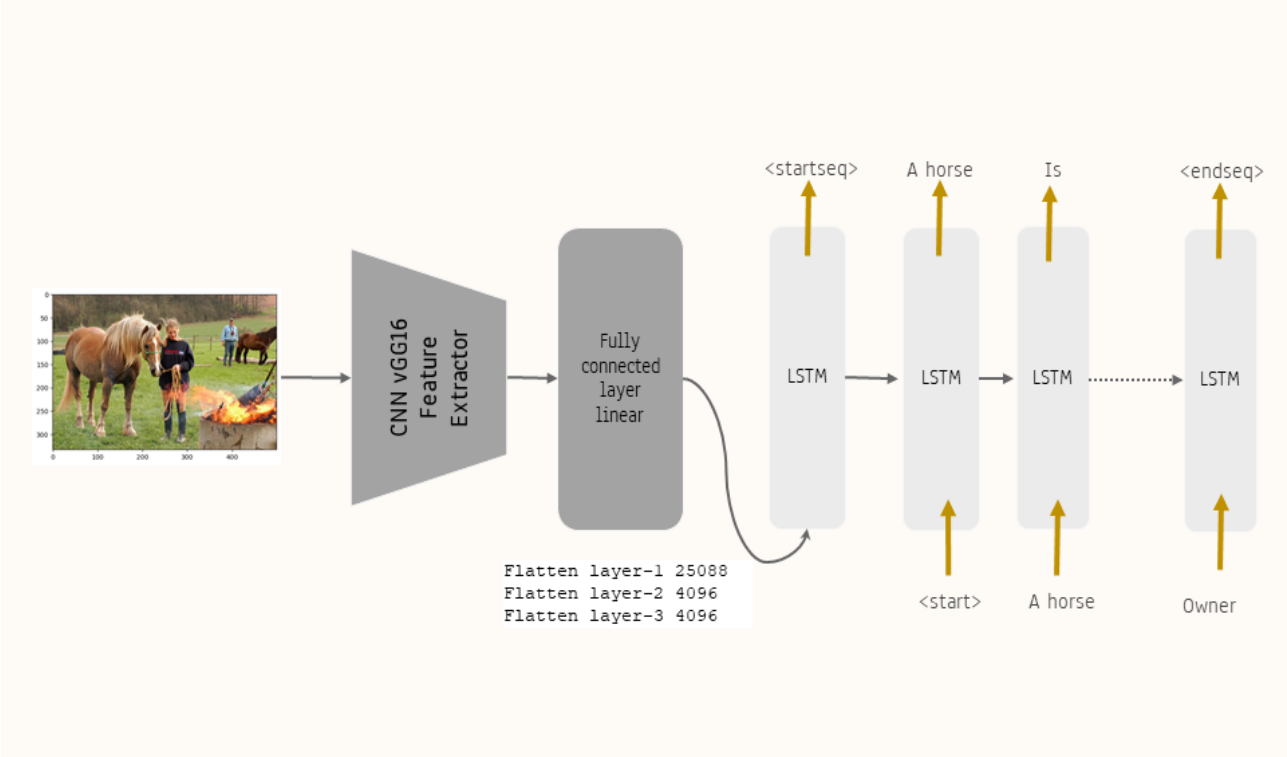

Figure 3: The basic architecture of our visual question generation (VQG) model

4.1.2 Architecture Overview:

The Visual Question Generation (VQG) framework employs an architecture that integrates an encoder built on the VGG16 neural network with a decoder that utilizes Long Short-Term Memory (LSTM) networks shown in figure - 4. The VGG16 encoder is responsible for distilling feature vectors from the visual inputs, which are subsequently stored in a pickle file for future retrieval. The accompanying dataset of questions undergoes a cleansing process to affix sequence terminators and eliminate characters that are not letters. These questions are then converted into numerical representations through a tokenization process, which standardizes their length, forming a comprehensive index of vocabulary. Utilizing these refined text sequences in tandem with the image-derived feature vectors, the LSTM decoder is trained to sequentially forecast subsequent words within a query sequence. It commences with an input layer accepting images of size 224x224 pixels with three color channels. The network consists of five blocks, each with convolutional layers using filters of increasing depth—from 64 in the first block to 512 in the last block—to extract features from the input image. Each block is followed by a max-pooling layer to reduce the spatial dimensions of the feature maps. After the fifth pooling layer, the network flattens the 3D feature maps into a 1D vector and passes it through two fully connected layers with 4096 units each. Overall, the model has approximately 134.26 million trainable parameters, occupying 512.16 megabytes of memory, which signifies its capacity to capture complex features and perform high-level image recognition tasks.

4.2 Technical Details of Visual Question Answering Model:

4.2.1 Dataset for VQA:

for a Visual Question and Answer Generation model, utilizing the CLEVR dataset, which contains around 850,000 questions and their corresponding answers. To simplify the process, a subset of 1,000 images, along with their questions and answers, was extracted from the dataset. These questions were then categorized into three distinct types: arithmetic questions that likely involve numerical reasoning about the visual content, binary questions that expect yes/no answers, and a third category labeled

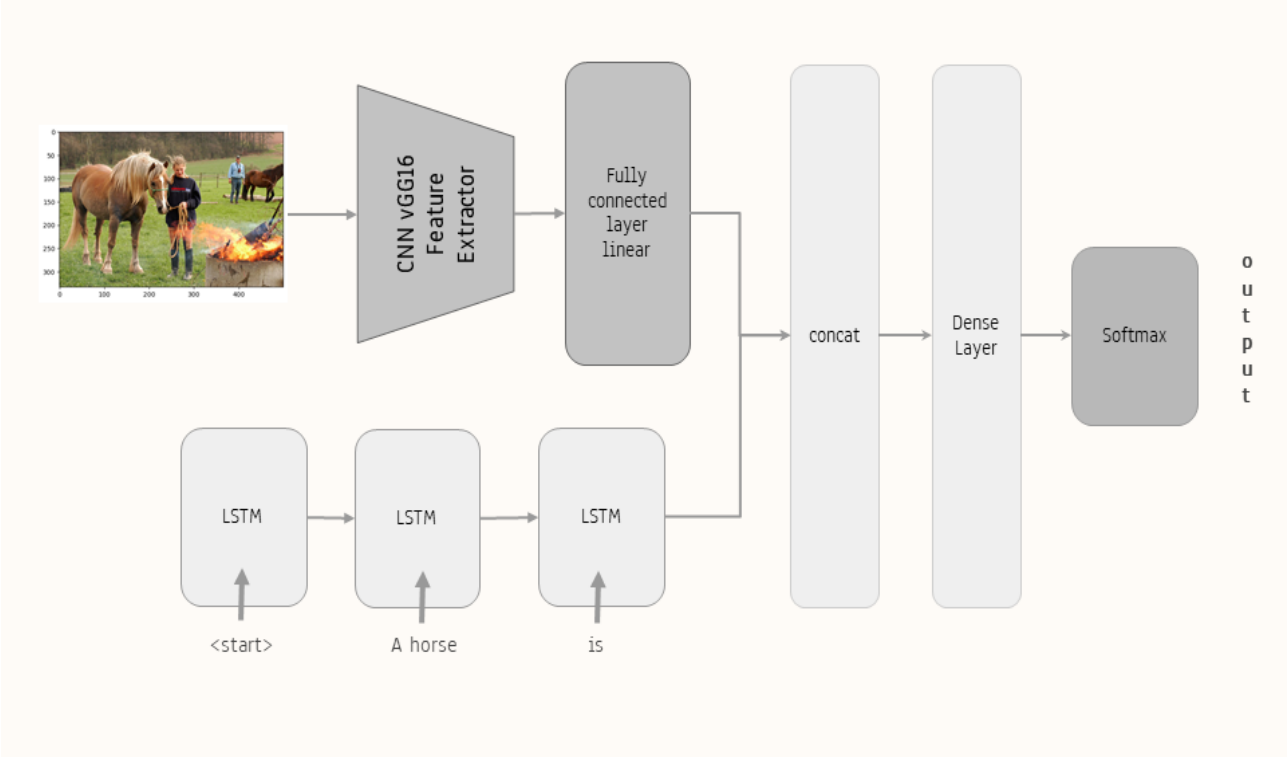

Figure 4: The basic architecture of our visual question answer (VQA) model

'other', which could encompass a variety of question types beyond the first two. This categorization is instrumental in creating a model capable of understanding and generating different types of questions based on visual data, enhancing its versatility and depth of cognitive processing.

4.2.2 Architecture Overview:

Visual Question Answering (VQA) model designed to interpret images and answer questions pertaining to them. Initially we prepare a vocabulary set for text processing. The model architecture integrates a MobileNetV2-based CNN for image feature extraction and a series of LSTM layers within an RNN to process encoded questions. These components are fused to predict answers, with the model trained on the dataset using a custom learning rate schedule and performance evaluated by comparing predicted answers to actual answers, demonstrating the model's learning and prediction capabilities. The model described is a sophisticated deep learning architecture designed for Visual Question Answering tasks, combining convolutional and recurrent neural networks. It starts with an input layer tailored for 200x200 pixel images, followed by a series of Conv2D and BatchNormalization layers, designed in the style of a MobileNetV2 architecture with depthwise separable convolutions for efficient feature extraction. This is coupled with a recurrent neural network comprising bidirectional LSTM layers, responsible for processing textual input. The model then merges these two streams—image features and processed text—using concatenation. The final output is obtained through a dense layer with 101 units, indicating the model is likely predicting answers from a fixed set of categories. With over 9 million trainable parameters, this model represents a complex system capable of understanding and answering questions about images, balancing the intricacies of both visual perception and language understanding.

5 Survey Survey System Design:

To explore the effectiveness of contextual visual reasoning in assisting individuals with visual impairments in performing reasoning tasks, we have created a web application utilizing HTML, CSS,

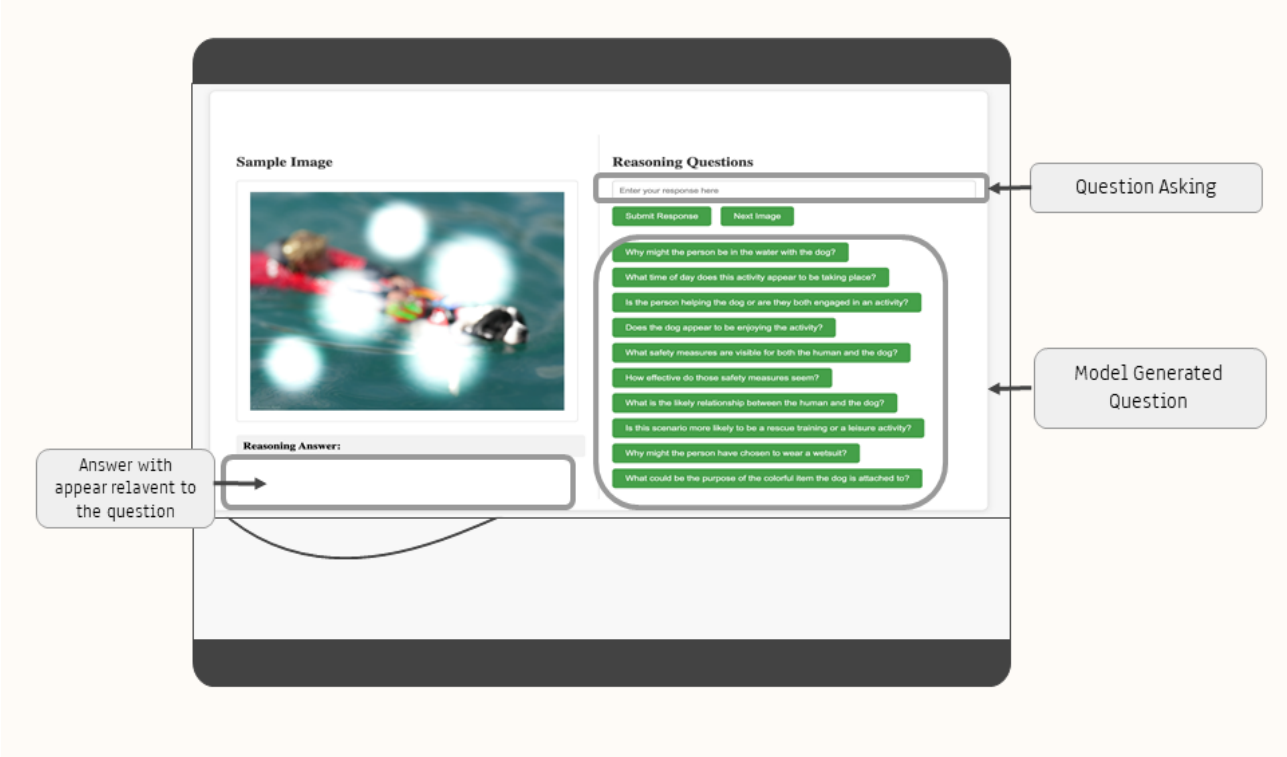

Figure 5: Front-end of the Visual Reasoning System

JavaScript, and Bootstrap for the front end, with Flask and Python powering the back end. Recognizing the limitation of not having direct access to visually impaired users, we have adapted our survey to simulate the diverse visual experiences encountered by people with visual impairments. This adaptation aims to ensure that the results of our study, both quantitative and qualitative, closely mirror those we would obtain from the target demographic. Our system employs a Visual Question Generation (VQG) model to produce questions. When a user selects a question, the query is sent to the back-end Visual Question Answering (VQA) model, which then generates a reasoning answer. This interactive feature serves to demonstrate the types of inquiries visually impaired users might pose to the VQA system, effectively initiating the conversation. Additionally, the top input box functions as a conversational agent, allowing users to input custom questions as desired show in figure - 5.

6 Result

6.1 Quantitative Result:

6.1.1 Visual Question Generation:

The Visual Question Generation (VQG) model's performance on the FLICKR 8K dataset was assessed using BLEU score metrics, which quantitatively evaluate the textual output quality on a scale from 0 to 100, with higher values indicating better quality. These scores typically assess machine translation quality by comparing the machine-generated text to a set of reference translations. The Google NIC [9] and Log Bilinear models [10] demonstrate diminishing scores from BLEU-1 to BLEU-4, reflecting reduced accuracy in generating longer text sequences shown on table-1. In contrast, the "Custom VQG Model" surpasses these models on the BLEU-1 and BLEU-2 metrics, indicating superior question generation when it comes to shorter sequences. The evaluation on BLEU-3 and BLEU-4 was not deemed necessary since the generated questions are often less than four words long. The enhanced performance of the "Custom VQA Model" is likely attributed to its training on five

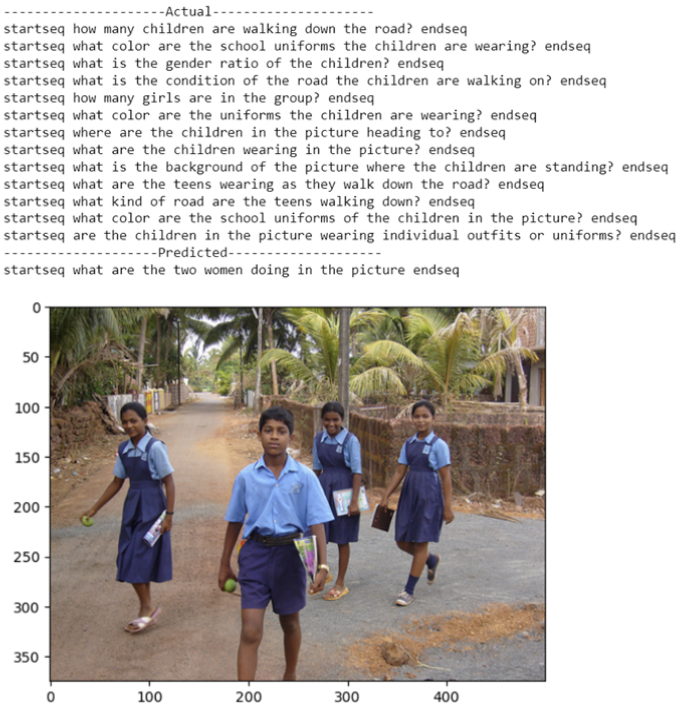

Figure 6: Sample Model Output for Visual Question Generation (VQG) model

questions per image compared to the single-caption approach of the Google NIC and Log Bilinear models, which may not provide as diverse a linguistic context for the model to learn from. A sample response is shown in the figure -6.

DATASET	**Model**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
FLICKR 8K	Google NIC [9]	63	41	27	-
	Log Bilinear [10]	65.6	42.4	27.7	17.7
FLICKR 8K (1000)	Custom VQG Model	75	57	-	-

Table 1: VQG comparison on the FLICKR 8K dataset with original paper with our custom model

6.1.2 Visual Question Answer:

Visual Question Answering (VQA) models were assessed on the CLEVR dataset, which comprises 850,000 questions. The VQA model developed by Stanislaw et al. achieves an overall accuracy of 52.64%, excelling in the Arithmetic Reasoning category with a performance of 75.55%. This performance suggests that the model is particularly proficient at processing questions that involve numerical and quantitative analysis. In comparison, the "Custom VQA model" registers a lower overall accuracy of 48.02% but exhibits a marginally better capacity in Binary Reasoning with an accuracy of 40.2%, denoting its effectiveness in answering questions that are binary in nature, such as yes/no or true/false. Nonetheless, it falls behind the Stanislaw et al. model in both the Arithmetic Reasoning and Other categories. The "Other" category includes a variety of reasoning question types not covered by the first two categories, with both models showing comparable performances. The red dot adjacent to the 75.55% in Arithmetic Reasoning could signify a noteworthy achievement or an important annotation concerning that result. The training, limited to only 2 epochs on the custom model, maybe the underlying reason for the lower accuracy of the "Custom VQA model." A sample response from VQA model is shown in figure - 7

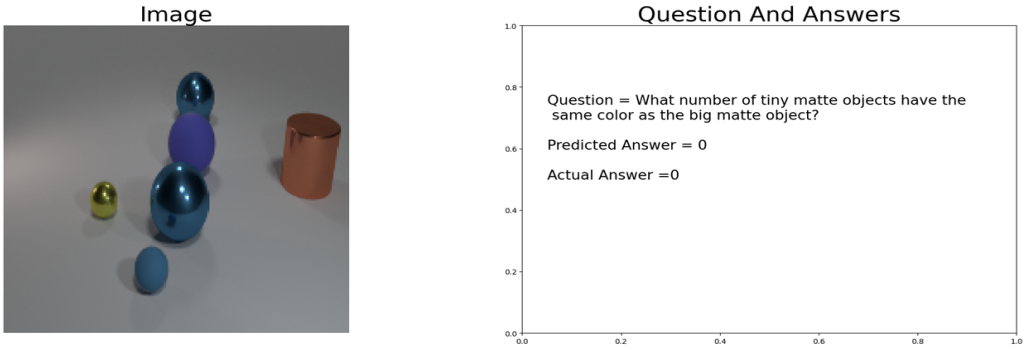

Figure 7: Sample Model Output for Visual Question Generation (VQG) model

DATASET	Model	Overall	Arithmetic	Binary	Other
CLEVR 850000	Stanislaw, et al [11]	52.64	75.55	33.67	37.37
	Custom VQA model	48.02	55.67	40.2	33.9

Table 2: Performance comparison of VQA models on the CLEVR dataset.

6.2 Qualitative Result:

In our study, we assessed the approach of contextual visual reasoning, where we evaluated this approach to enhance visually impaired people's understanding of visualizations. As our initial VQG and VQA models were not fully efficient, we integrated the Vision-GPT API for the study's effectiveness. We modified images to simulate the visual experience of visually impaired individuals. This enabled us to test our system with regular users. The survey involved 8 participants from CIS-5603, each session lasting 30 minutes and consisting of two parts. In the first part of the survey, users were presented with seven images, each paired with a question and an answer. Subsequently, they answered interview questions regarding the images' relevance, question variety, personal thoughts, answer correctness, clarity, and comprehension.

The survey aimed to collect feedback on the relevance and accuracy of the system-generated questions and answers, system usability including the text box feature, and its effectiveness in enhancing users' understanding of the images. Participants were also asked to identify potential improvements for visually impaired users and suggest enhancements to the question-answer experience within the app. All 8 participants provided positive feedback on the system's usefulness in better understanding images. The detailed survey results are shown in figure -**??** and explained as follows:

- **Questions Relevant:** Median $(u) = 4.5$, Upper Whisker $(u+\sigma) = 5.0$, Lower Whisker $(u-\sigma) = 4.0$. This indicates the model's relevance in questioning.
- **Own Thinking:** Median $(u) = 2.5$, Upper Whisker $(u+\sigma) = 3.0$, Lower Whisker $(u-\sigma)$ just above 1.0. This shows alignment with the thought process of visually impaired users.
- **Utilize Text for Feature:** Median $(u) = 3.0$, Upper Whisker $(u+\sigma)$ close to 4.0, Lower Whisker $(u-\sigma)$ slightly above 2.0. This suggests a reduced need for user-initiated questions.
- **Variety Aspects:** Median $(u) = 4.0$, Upper Whisker $(u+\sigma)$ just below 5.0, Lower Whisker $(u-\sigma)$ near 3.5. This confirms diverse aspect coverage in questioning.

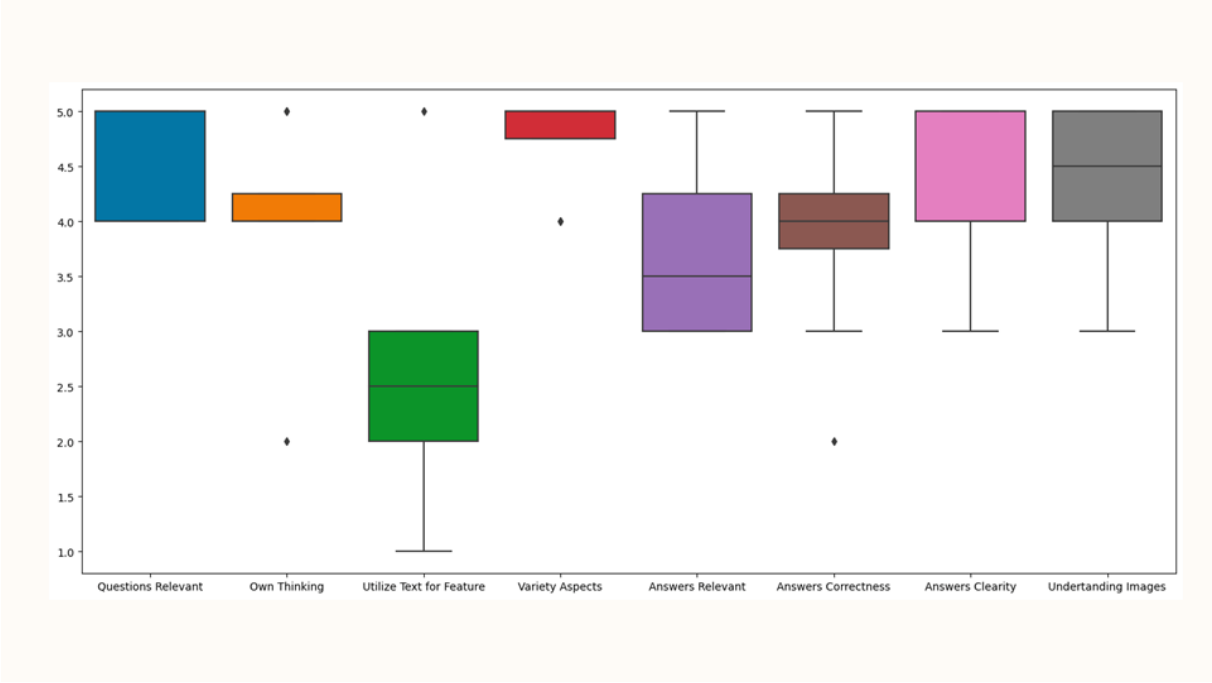

Figure 8: Survey result based on different interview questions

- **Answers Relevant:** Median (u) = 4.0, Upper Whisker ($u+\sigma$) just below 5.0, Lower Whisker ($u-\sigma$) approximately 3.5.
- **Answers Correctness:** Median (u) = 4.0, Upper Whisker ($u+\sigma$) slightly below 5.0, Lower Whisker ($u-\sigma$) roughly 3.5. This indicates that most responses are correct.
- **Answers Clarity:** Median (u) = 4.0, Upper Whisker ($u+\sigma$) near 5.0, Lower Whisker ($u-\sigma$) approximately 3.5. This reflects clarity in responses.
- **Understanding Images:** Median (u) = 4.0, Upper Whisker ($u+\sigma$) just below 5.0, Lower Whisker ($u-\sigma$) close to 3.5. This shows enhanced image comprehension for visually impaired people.

Generally, the majority of participants believe that this method significantly enhances understanding of visualizations, which is particularly beneficial for people with visual impairments. Participant feedback included the following comments:

> "I am impressed by the application's functionality; it has enabled me to comprehend aspects of images that were not initially apparent to me, effectively allowing me to 'visualize' them."

Participants offered insights on the interactive aspect of engaging in dialogue with images, pinpointing areas that require enhancement. These include the necessity for more succinct text tailored for visually impaired users and the proposal to incorporate a voice-activated question-and-answer system in subsequent studies.

> "The ability to interact with an image through conversation is quite fascinating. However, there are times when the responses provided by the system do not correspond to the visual content of the image."

7 Discussion

The findings from our user study indicate that the visual reasoning approach offers a detailed perspective on entire images. This method is particularly beneficial for visually impaired individuals, as it employs contextual reasoning to provide initial questions, aiding in deeper exploration and understanding of the images. This approach stands out from existing methods like conversational agents, OCR generators, or image summaries. Nonetheless, the system has several limitations. One major concern is the potential for inaccurate answers from the model, leading to misinterpretation. Additionally, since the system currently relies on text-based questions, and considering the challenges visually impaired people face with text, integrating a voice-based system could be advantageous. In terms of our VQG and VQA models, their accuracy is not yet sufficient for real-world application. Therefore, using high-quality data and improving model accuracy is essential for future applications. Overall, this approach shows promise for assisting visually impaired users.

Acknowledgement

I would like to express my sincere thanks to my classmates who generously volunteered for this study. The theoretical foundations were gathered from analyticbidya.com [12], and the modifications to the VQA models were based on open-source code from a repository [13].

Attachment

Along with this project report, python code for VQG and VQA model, flask code for the system, demo video of the system, survey result and presentation slide are included.

References

[1] https://www.who.int/health-topics/blindness-and-vision-loss

[2] http://surl.li/oktua

[3] Kuriakose, Bineeth, Raju Shrestha, and Frode Eika Sandnes. "Tools and technologies for blind and visually impaired navigation support: a review." IETE Technical Review 39.1 (2022): 3-18.

[4] Al-Muqbali, Fatma, et al. "Smart Technologies for Visually Impaired: Assisting and conquering infirmity of blind people using AI Technologies." 2020 12th Annual Undergraduate Research Conference on Applied Computing (URC). IEEE, 2020.

[5] Felix, Shubham Melvin, Sumer Kumar, and A. Veeramuthu. "A smart personal AI assistant for visually impaired people." 2018 2nd international conference on trends in electronics and informatics (ICOEI). IEEE, 2018.

[6] Kim, Dae Hyun, Enamul Hoque, and Maneesh Agrawala. "Answering questions about charts and generating visual explanations." Proceedings of the 2020 CHI conference on human factors in computing systems. 2020.

[7] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.

[8] Fernandes, Daniel Louzada, et al. "Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs." arXiv preprint arXiv:2202.05331 (2022).

[9] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).

[10] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.

[11] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

[12] https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/

[13] https://www.kaggle.com/code/marcelosabaris/visualquestionanswering