Project Proposal

Smart VQA for Visual Assistance

Group Number: 6

Group Member Names:
Deep Patoliya
Kruti Kotadia
Kunal Jain

Objective:

The goal of this project is to develop a deep learning model that can perform Visual Question Answering (VQA) using the VizWiz dataset, which contains real-world images and questions posed by visually impaired individuals. The model will integrate both visual and textual data to answer questions about images accurately. This project aims to enhance accessibility for the visually impaired by building an intelligent system that can assist in understanding and interpreting visual content.

Data Set Description:

Overview: The VizWiz VQA dataset contains visual questions posed by people who are blind, providing real-world images and corresponding questions. The data focuses on accessibility challenges, with some questions marked as unanswerable due to poor image quality or ambiguous content.

- **Planned Size for Colab**: Approximately 10,000 images and their associated questions and answers will be used, ensuring the subset fits within Colab's resource limits.

- **Number of Rows/Columns**: Each data point includes an image, a question, and a corresponding answer (or "unanswerable" flag).

Sample Predictors:

- Image features extracted via a pre-trained CNN.
- Text-based question embeddings (e.g., via LSTM or Transformer).

Dataset Link: https://vizwiz.org/tasks-and-datasets/vqa/

Interesting Aspect: Many questions are unanswerable, presenting a unique challenge of identifying when the visual content is insufficient for answering.

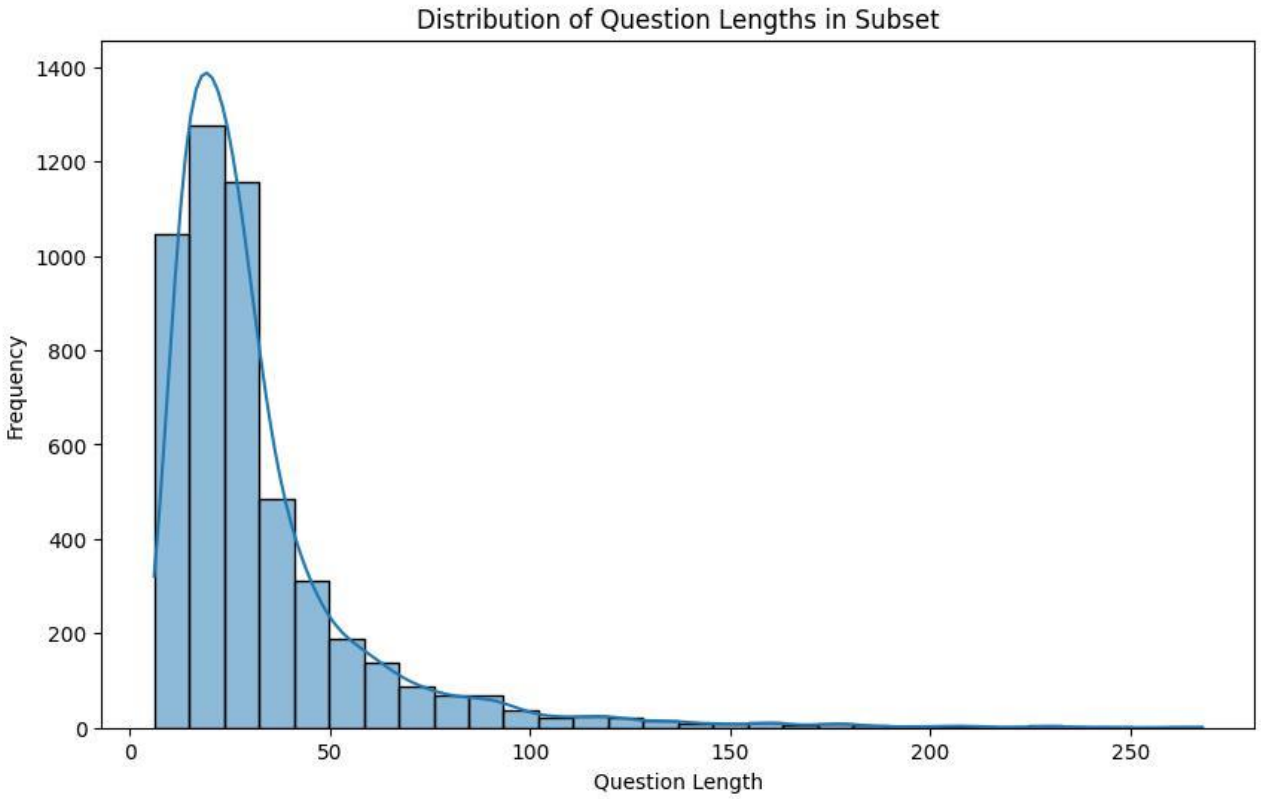
Preliminary Data Exploration:

Summary Statistics:

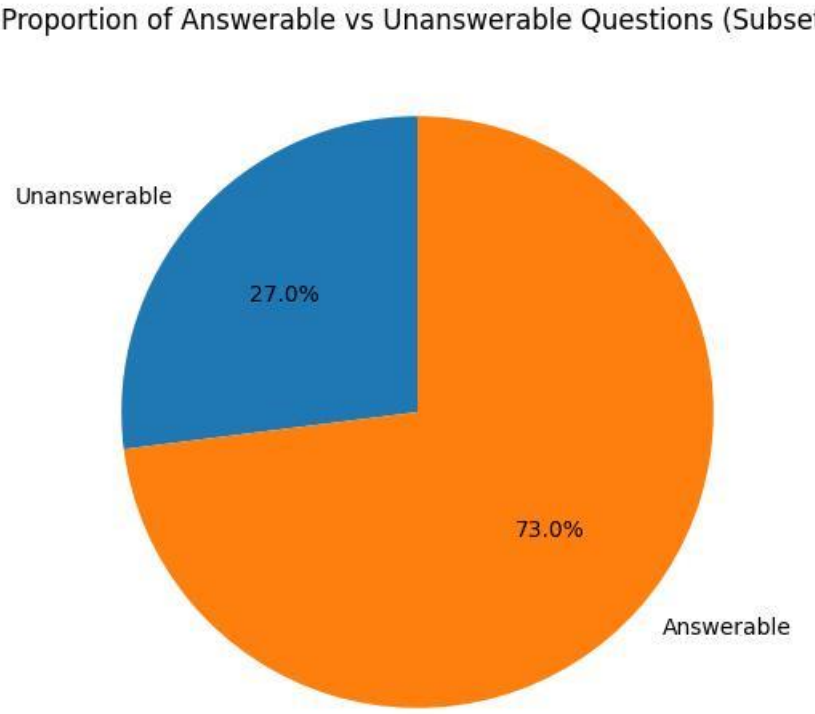
- For this project, we are using a subset of **10,000 images** and their corresponding questions and answers from the VizWiz VQA dataset. Approximately **20%** of the questions are unanswerable, highlighting the dataset's unique challenge. The average question length is around **6–12 words**, and the top 10 most common words include terms like "this," "color," and "kind," indicating a focus on identifying objects and attributes in the images. The data includes a wide range of image qualities, further complicating the question-answering task.

Visualizations:

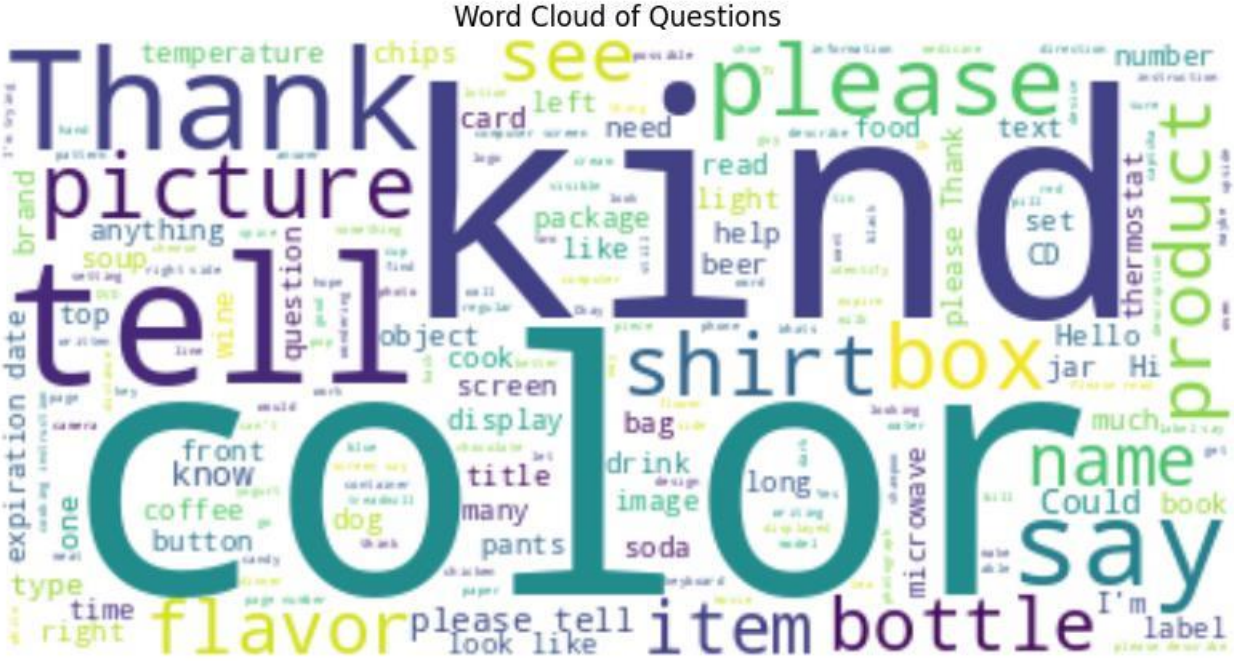
- **Distribution of Question Lengths**: The histogram shows that most questions are relatively short, with the majority containing fewer than 20 words. This suggests that users tend to ask concise questions.



- **Pie Chart of Answerable vs Unanswerable Questions**: The pie chart shows that 73% of the questions are answerable, while 27% are unanswerable, highlighting the challenge of dealing with ambiguous or poor-quality images.



- **Word Cloud of Questions**: This word cloud shows the most frequently occurring words in the questions, with terms like "kind," "color," and "tell" appearing prominently. This indicates that many questions focus on object identification and description.



Proposed Data Exploration:

1. **Relationship Between Question Length and Answerability**: Investigate whether longer questions are more likely to be answerable or unanswerable.

2. **Common Object Types in Answerable Questions**: Analyze the most frequent object categories present in answerable questions.

3. **Impact of Image Quality on Answerability**: Explore if lower-quality images are more likely to have unanswerable questions.

Proposed Predictions:

we are planning to get the following predictions:

- **Answerability Detection**: The model will predict whether a question is answerable based on the image and question text.

- **Answer Prediction**: The model will predict the correct answer to the question using attention mechanisms to focus on relevant parts of the image.

- **Impact of Image Quality on Answerability**: The model will predict if image quality affects the likelihood of an unanswerable question.

Note: We will use only a subset of the dataset due to the computational limits of Google Colab. However, this will not affect the quality of the results, as we plan to fine-tune pre-trained models, which ensures efficient learning even with a reduced dataset.