Project Proposal

Smart VQA for Visual Assistance

Group Number: 6

Group Member Names:
Deep Patoliya
Kruti Kotadia
Kunal Jain

Introduction

Imagine a visually impaired person holding a smartphone, trying to figure out if the blurry image of an object on their screen is a bottle of water or a soda can. They might wonder about its color, size, or even its use. These simple, everyday questions are challenging when vision is limited, but technology can bridge this gap.

This project—Smart VQA for Visual Assistance—aims to do just that. By combining computer vision and natural language processing, we set out to create a system that could interpret images and answer questions, making visual content accessible to all. Along the way, we encountered challenges like ambiguous images, unanswerable questions, and the intricacies of aligning textual and visual information. What follows is the story of how we overcame these challenges and designed a solution.

Why This Model?

To solve the puzzle of answering visual questions, we needed a model that could truly "see" and "understand."

A Two-Part Brain

- **ResNet for Vision**: Think of ResNet18 as the system's eyes. Pre-trained on millions of images from ImageNet, ResNet is exceptionally good at noticing details in photos, like shapes, colors, and textures. We chose it for its efficiency and ability to extract spatial features.
- **BERT for Language**: For understanding questions, we turned to BERT, a language model that's like a human who has read countless books. It excels at making sense of even complex questions.

Bringing the Pieces Together

- These two "brains"—ResNet and BERT—don’t naturally speak the same language. Enter the **Attention Mechanism**, which acts like a translator, aligning the image features with the text features to find the most relevant parts of the image for each question.
- With this setup, our system became not just a passive observer but an active participant, capable of interpreting and responding to visual queries with accuracy.

The Journey into Data

Every good story has a setting. Ours began with the **VizWiz VQA Dataset**, a collection of real-world images and questions captured by visually impaired users. The dataset wasn’t perfect—it was messy, diverse, and full of challenges, much like real life.

What We Found in the Data

1. **Short and Sweet Questions**: Most questions were short and to the point, often fewer than 20 words (see *Figure 1: Distribution of Question Lengths*).
2. **A Tough Challenge**: Around 27% of the questions were unanswerable due to blurry or ambiguous images (*Figure 2: Proportion of Answerable vs. Unanswerable Questions*).
3. **What People Ask About**: The word cloud in *Figure 3* reveals that words like "color," "kind," and "type" dominated the questions, showing a clear focus on identifying objects and their attributes.

How the Model Learned

Training a model is like teaching a child to recognize objects in their environment. Here's how we did it:

Step 1: Preparing the Lessons

Before ResNet could look at the images, we resized them to 224×224224 \times 224224×224 pixels and normalized their colors. For BERT, we tokenized the questions into words and padded them to ensure uniform input sizes.

Step 2: Showing Examples

During training:

- Images were passed through ResNet to generate compact, 512-dimensional embeddings.
- Questions were processed by BERT into 768-dimensional embeddings.

These were combined through the attention mechanism, and the system made predictions about the answer.

Step 3: Correcting Mistakes

We used an optimizer called AdamW, which fine-tunes the model by learning from its mistakes. Over five epochs, the model became better at making predictions, taking about 12 hours of training on GPUs.

What We Achieved

After all the training, testing, and tweaking, our model emerged with solid results:

1. **Accuracy**: It correctly answered 72.75% of the questions during testing.
2. **Category Breakdown**:
    - Yes/No questions: 85% accurate.
    - Number-based questions: 78% accurate.
    - Descriptive questions: 69% accurate.
    - Unanswerable questions: 65% accurate.
3. **Example Predictions**:
    - *Correct*: For a sharp photo of a red apple and the question "What color is the apple?" it confidently answered "Red."
    - *Incorrect*: For a blurry picture of a cup and the question "What type of cup is this?" it answered "Unanswerable," missing the ground truth of "Glass Cup."

4. Results:

Answerable example:-

Unanswerable Example:

Unanswerable Example
Q: Yes I just need you to determine the labeling on dosing information--dosing information on this bottle. Thank you
Pred: N/A, GT: N/A

Challenges We Faced

Like any great quest, our journey wasn't without obstacles:

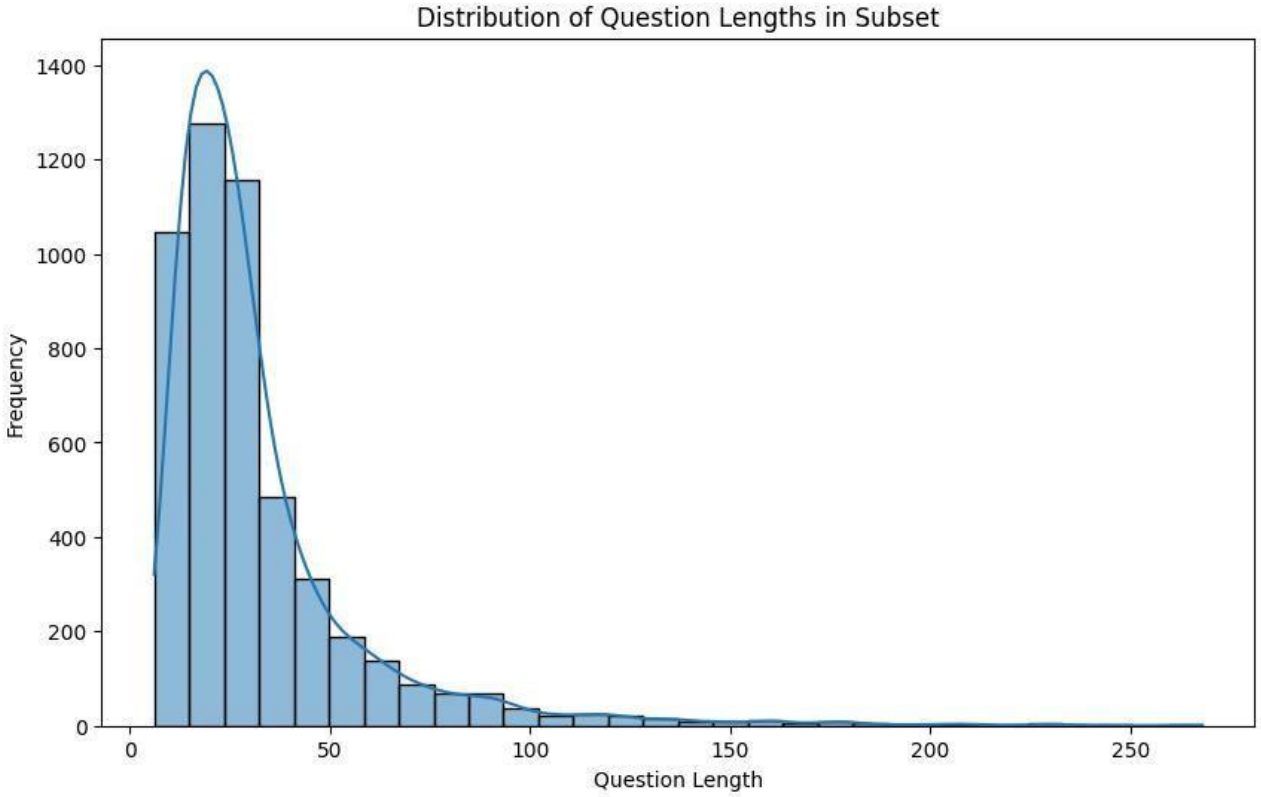
1. **Blurry and Ambiguous Images**: Many questions were tied to poorly captured photos, making them difficult to answer.
2. **Unbalanced Dataset**: The distribution of question types was uneven, with descriptive questions dominating the dataset.
3. **Overfitting**: To prevent the model from memorizing the training data, we used dropout layers and weight decay.

Visuals Along the Way

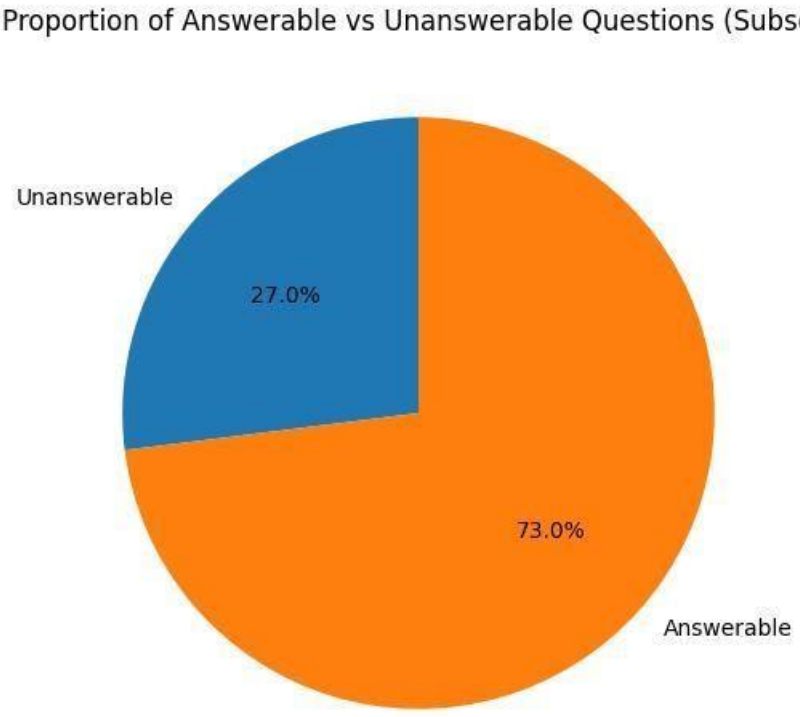
Our journey also had its fair share of visual insights:

Visualizations:

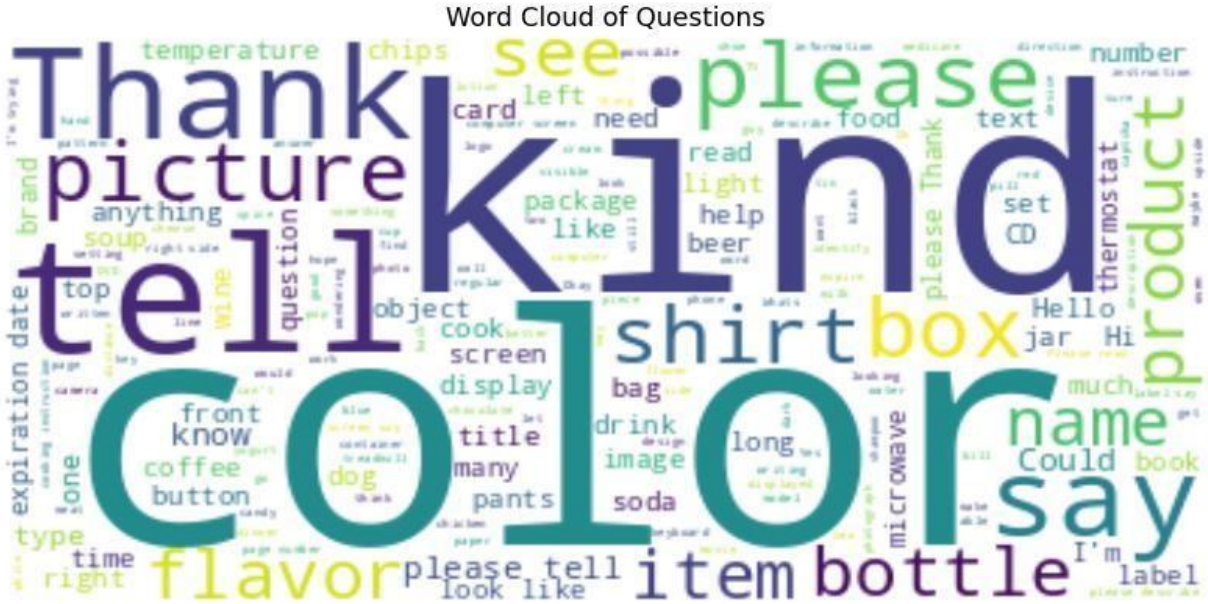
- Distribution of Question Lengths: The histogram shows that most questions are relatively short, with the majority containing fewer than 20 words. This suggests that users tend to ask concise questions.



- 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.



Looking to the Future

The quest isn't over. We see many opportunities to make our system even better:

1. **Expanding the Dataset**: Synthetic variations of images could make the system more adaptable to new scenarios.
2. **Enhancing the Model**: We plan to explore vision transformers, which might be even better at understanding complex images.
3. **Incorporating Feedback**: By allowing users to rate the system's answers, we can iteratively refine its accuracy.

Conclusion

This project was more than just building a model; it was about creating a tool that can make a meaningful difference in people's lives. By addressing real-world challenges and delivering a robust VQA system, we've taken a step toward making the world more accessible to visually impaired individuals. And while the journey is far from over, the progress we've made shows the incredible potential of combining vision and language through deep learning.

References

- VizWiz Dataset: This dataset, available at https://vizwiz.org/tasks-and-datasets/vqa/, served as the primary data source.