Tackling Ambiguity in Visual Question Answering: A Clarification Feature with BLIP-2

Moe Ito
Mount Holyoke College
ito23m@mtholyoke.edu

***Abstract*— This paper introduces a clarification feature to address the challenge of ambiguity in questions posed to Visual Question Answering (VQA) systems. Ambiguity often arises from vague pronouns (e.g., ”it,” ”this,” ”that”) or general terms (e.g., ”thing,” ”object”) without sufficient context. The proposed feature detects such ambiguities and prompts users to refine their questions, enhancing response accuracy. The implementation employs BLIP-2, a state-of-the-art vision-language model, and utilizes the VQA v2.0 dataset. A prototype system was developed and demonstrated with a specific image, successfully identifying and resolving ambiguous user questions. The results highlight the feasibility of incorporating a clarification mechanism to improve VQA systems. Future work will explore broader applications and more robust evaluation methods.**

***Index Terms*— Visual Question Answering; Natural Language Processing; Ambiguity; Clarification; BLIP-2**

I. Topic Introduction

Visual Question Answering (VQA) is a task in computer vision that combines visual recognition and natural language processing. It involves generating a natural language answer to a question posed about a given image. This task is significant for its potential to bridge the gap between visual perception and language understanding, addressing both high-level reasoning and detailed visual comprehension.

VQA gained significant attention with the publication of the paper ”VQA: Visual Question Answering” by Antol et al. in 2015 [1], which introduced a standardized dataset and benchmark for the task, and with the subsequent competition held at CVPR 2016. Together, these developments established VQA as a prominent research area in computer vision, enabling researchers to evaluate models consistently and driving rapid advancements in the field. The potential applications of VQA are broad and impactful. It can assist visually impaired individuals by interpreting their surroundings, aid medical professionals in analyzing diagnostic images, and streamline tasks like content retrieval and safety inspections. For example, VQA systems could help identify hazards in industrial settings or enable users to locate specific scenes in videos efficiently.

To evaluate progress in VQA, datasets play a crucial role. One widely used benchmark is the VQA v2.0 dataset, proposed by Goyal et al. [2]. This dataset improves upon the original VQA dataset by addressing answer imbalance and doubling the number of image-question pairs to approximately 1.1 million. It pairs each question with two similar images that lead to different answers, ensuring robustness in model evaluation. For example, in Figure 1, the question ”Who is wearing glasses?” is associated with two images: one where a man is wearing glasses and another where a woman is wearing glasses. Both images include both a man and a woman, highlighting how the dataset carefully balances image-question-answer relationships. Figure 1 shows sample entries from the VQA v2.0 dataset, illustrating its open-ended question-and-answer pairs and its design to accommodate diverse visual contexts.

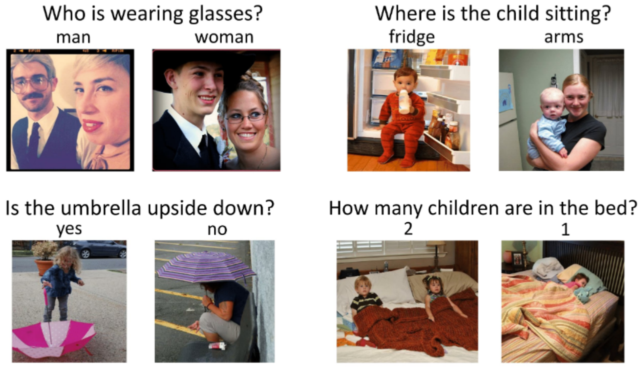

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Fig. 1. Examples from the VQA v2.0 dataset.

II. Existing Methods

Traditional Approaches: Early Visual Question Answering (VQA) models relied on Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory networks (LSTMs) for processing question text. Antol et al. (2015) [1] demonstrated these methods on the original VQA dataset, using pre-trained VGGNet and simple question encoding techniques. While effective for straightforward queries, such as object recognition, these methods struggled with more complex questions requiring deeper reasoning, such as counting or spatial understanding. These limitations highlighted the need for more sophisticated mechanisms to align visual and textual information.

Attention Mechanisms and Transformers in VQA: Attention mechanisms laid the foundation for significant advancements in VQA. Lu et al. (2016) [3] introduced a Hierarchical Co-Attention Model that applied attention jointly to image regions and question words, capturing fine-grained alignments. This approach enhanced the alignment between visual and textual features, improving model accu-

TABLE I
SUMMARY OF EXISTING METHODS IN VQA

Method	Key Features	Strengths	Limitations
Traditional Approaches	CNNs for image features, LSTMs for question processing	Effective for simple VQA tasks	Struggles with complex queries
Co-Attention Models	Joint attention to image and question	Captures modality alignments	Computationally expensive
Transformer-Based Models	Self-attention mechanisms for vision and language	Accurate contextual understanding	High computational cost
BLIP	Pre-trained encoders for multimodal tasks	Versatile for VQA and captioning	Limited handling of ambiguous queries
BLIP-2	Combines frozen encoders with LLMs	High efficiency and scalability	Requires further tuning for ambiguity

racy on complex VQA tasks. Building on the success of attention mechanisms, the Transformer architecture introduced in ”Attention Is All You Need” (2017) revolutionized the field of machine learning. Transformer-based models such as ViLBERT [4] and UNITER [5] have significantly advanced VQA by enabling seamless integration of visual and textual inputs. These models utilize self-attention and co-attention mechanisms to align and interpret complex relationships within image-question pairs, setting new benchmarks in VQA accuracy.

BLIP and BLIP-2: BLIP and BLIP-2 are among the recently proposed models. BLIP (Bootstrapping Language-Image Pre-training), introduced by Salesforce in 2022 [6], serves as a unified framework for vision-language tasks such as image captioning and VQA. It leverages pre-trained encoders to align visual and textual modalities, enabling robust cross-modal reasoning. BLIP-2, a subsequent advancement proposed in 2023 [7], extends this framework by integrating frozen pre-trained vision models with large language models (LLMs). To bridge the gap between these two modalities, BLIP-2 employs a Querying Transformer (Q-Former), which facilitates efficient alignment while minimizing the number of trainable parameters. Despite its lightweight design, BLIP-2 achieves state-of-the-art performance across various vision-language tasks, demonstrating enhanced scalability and computational efficiency.

Table I summarizes key existing methods in VQA, highlighting their strengths and limitations. These methods demonstrate the evolution of techniques from traditional approaches to recent advances.

III. Open Challenges

A. Current Limitations

While existing VQA datasets and models assume that questions are clearly defined, real-world interactions often involve ambiguous questions with vague pronouns (e.g., ”it,” ”this”) or omitted context. This ambiguity hinders models from providing definitive answers, highlighting a key limitation in their ability to handle less structured, real-world queries [8].

In related fields like Knowledge-Based Question Answering (KBQA), mechanisms for generating clarification questions have been explored to address similar challenges [9]. However, such methods are still underexplored in VQA, where ambiguity in visual and textual contexts presents unique challenges. This gap underscores the need for novel approaches to detect and resolve ambiguities in VQA, enabling more practical and real-world applications.

B. Research Idea

To address the challenges inherent in existing VQA systems discussed in the previous section, this research aims to propose a VQA system equipped with a clarification feature. The following are the core components of this proposed idea:

1) **Ambiguity Detection**: The proposed system leverages BLIP-2’s pre-trained capabilities in both natural language and visual understanding to identify user queries that lack sufficient specificity. The system will automatically detect ambiguous expressions, such as vague pronouns (e.g., ”it,” ”this,” ”that”) and general terms (e.g., ”thing,” ”object”), by analyzing the interplay between the image context and the accompanying textual query. This enables the system to identify when a question requires additional clarification before an accurate answer can be generated.
2) **Clarification Question Generation**: Upon detecting ambiguity, the system generates follow-up questions to clarify the user’s intent. BLIP-2, fine-tuned with a custom dataset containing ambiguous queries and their corresponding clarification questions, is used to dynamically formulate questions that address the detected ambiguity. For example, if the user asks, ”What is it?” in the context of an image, the system might respond, ”Are you referring to the object on the sofa or the one near the table?” By leveraging BLIP-2’s ability to align visual and linguistic features, the system ensures that the generated questions are both context-aware and precise.

IV. Concept to Code

The runnable source code associated with this proposed research idea is available on GitHub and can be found here: `https://github.com/MHC-FA24-CS341CV/beyond-the-pixels-emerging-computer-`

vision-research-topics-fa24/tree/main/code/14-vqa.

A. Dataset Description

This program uses the VQA v2.0 dataset [2], which consists of:

- **Questions**: Open-ended questions about images that require an understanding of vision, language, and commonsense knowledge to answer.
- **Annotations**: Human-annotated answers to each question, including multiple acceptable responses.
- **Images**: Real-world images sourced from the MS COCO dataset, covering diverse scenes and objects.

This program specifically utilizes the validation set of the VQA v2.0 dataset, reflecting the project's focus on demonstrating inference capabilities of the pre-trained BLIP-2 model. The smaller size of the validation set further supports streamlined demonstrations, making it a practical choice for this purpose.

B. Run Instructions and Connection to Research Idea

To execute the program, follow these steps:

1) Access the GitHub folder for this program via the link provided at the start of this section.
2) Download the following three required files from the link detailed in the README.md:
   - v2_OpenEnded_mscoco_val2014_questions.json
   - v2_mscoco_val2014_annotations.json
   - val2014.zip
3) Upload these files to your Google Drive.
4) Open the .ipynb file in Google Colab by clicking the **"Open in Colab"** button displayed at the top of the file.
5) Modify the file paths in the notebook to reflect the location of the dataset files in your Google Drive.
6) Run each cell in the notebook sequentially to observe the system's functionality and interact with the clarification feature.

This program connects directly to the proposed research idea by demonstrating the clarification feature in action. Specifically, it shows how ambiguous user queries are detected and resolved interactively, aligning with the research objective of enhancing VQA systems' contextual understanding and accuracy.

C. Results

To demonstrate the functionality of the system, a specific image from the validation set of the VQA v2.0 dataset was used. Figure 2 shows the image we used, along with the interactive VQA process shown in Figure 3. This process demonstrates how the system identifies ambiguous questions, prompts the user for clarification, and generates a final answer.

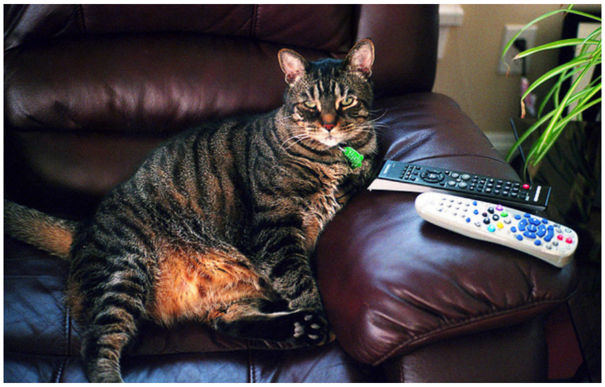
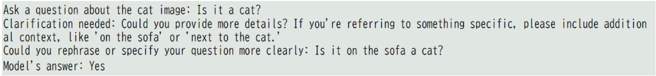


Fig. 2. The cat image used in the program.



```
Ask a question about the cat image: Is it a cat?
Clarification needed: Could you provide more details? If you're referring to something specific, please include addition
al context, like 'on the sofa' or 'next to the cat.'
Could you rephrase or specify your question more clearly: Is it on the sofa a cat?
Model's answer: Yes
```

Fig. 3. Output example of the interactive VQA process with clarification feature.

D. Future Work and Technical Implementations

The current implementation serves as a simplified demonstration, focusing on a specific image and predefined ambiguous terms to test the concept of ambiguity detection and clarification in VQA. However, scaling this idea to a robust and user-friendly system capable of handling diverse images and user-generated questions involves addressing several technical challenges:

- **Automated Ambiguity Detection**: The current implementation relies on manually defined ambiguous terms and context phrases. Future work involves fine-tuning BLIP-2 on a custom dataset containing diverse images and questions labeled with ambiguity annotations, enabling automatic detection of vague queries.
- **Scalability to Diverse Inputs**: While the current implementation is limited to a single image, the ideal system would allow users to upload arbitrary images and ask open-ended questions. Extending the system to dynamically handle diverse visual and textual contexts is crucial for broader applicability.
- **Dataset Expansion**: A large-scale dataset that includes ambiguous questions, clarified versions, and corresponding answers across various scenes and objects is essential. Such a dataset would facilitate fine-tuning and improve the system's ability to generalize to real-world scenarios.

These advancements aim to enhance the scalability and robustness of the system, enabling it to handle diverse images and questions effectively while maintaining high interpretive accuracy.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[2] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/9dcb88e0137649590b755372b040afad-Paper.pdf

[4] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[5] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[7] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[8] S. Kottur, J. M. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 153–169.

[9] J. Xu, Y. Wang, D. Tang, N. Duan, P. Yang, Q. Zeng, M. Zhou, and X. Sun, "Asking clarification questions in knowledge-based question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1618–1629.