Research Proposal

Visual Question Answering
based on Visual Programming Framework

Hao Xu xuhao57@mail2.sysu.edu.cn Sun Yat-sen University

December 21, 2023

Contents

1 Introduction

Visual Question Answering (VQA) presents a fundamental challenge at the intersection of computer vision and natural language processing. Its goal is to bridge the gap between visual perception and natural language understanding. The proliferation of visual data across various domains, including social media, surveillance, and e-commerce, has intensified the interest in VQA due to its potential applications in image search engines, virtual assistants, and recommendation systems. The task entails generating accurate and meaningful answers to questions related to visual content, such as images or videos.

Notably, current VQA methods [1] have made significant strides by leveraging deep learning techniques, including attention-based models, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) coupled with long short-term memory (LSTM) networks. These approaches have successfully tackled various challenges in VQA, resulting in state-of-the-art performance in terms of accuracy. However, certain limitations persist, curbing their practical utility and hindering their performance in real-world scenarios.

One of the primary shortcomings of existing methods pertains to their limited interpretability. Although these models yield precise predictions, they often operate as black boxes, rendering it arduous to comprehend the underlying decision-making process. This lack of interpretability poses concerns, particularly in critical applications where transparency and explainability are imperative.

Furthermore, effectively handling complex queries that necessitate context understanding remains a formidable challenge for existing VQA methods. Questions reliant on contextual knowledge or fine-grained visual details often yield suboptimal or erroneous answers. Contextual understanding proves vital for VQA systems to grasp relationships among objects, scenes, and actions, empowering them to reason and provide accurate responses. Unfortunately, current methods struggle to capture and exploit such contextual cues effectively.

Additionally, biases entrenched within the training data can lead to biased answers. Existing VQA datasets frequently skew towards particular classes or demographics, resulting in skewed distributions and unfair representation. It is crucial to address such bias to ensure fairness and inclusivity within VQA systems, particularly when deploying them in applications serving diverse user populations.

To overcome these limitations, this research notices a novel visual programming framework [5] designed explicitly for managing VQA tasks. This framework aims to merge visual data integration, programming logic representation, and advanced model integration to transcend existing challenges and limitations. By harnessing the potential of visual programming, which offers an intuitive and flexible interface to represent the requisite programming logic for query answering, the proposed framework strives to enhance the accuracy, efficiency, and overall usability of VQA systems.

This research plan encompasses a comprehensive exploration of existing VQA methods, analyzing their strengths, limitations, and areas necessitating improvement. Subsequent sections will elaborate on the proposed visual programming framework in detail, emphasizing its key components and elucidating how they effectively address the identified challenges associated with current VQA approaches. Furthermore, forthcoming enhancements and research direc-

tions will be discussed, enabling a continuous refinement of VQA systems.

In summary, the primary objective of this research plan is to introduce a novel visual programming framework designed for managing VQA tasks. It aims to address the limitations encountered by existing methods by integrating visual data, programming logic, and question processing capabilities. The framework seeks to enhance the accuracy and efficiency of VQA systems while ensuring transparency, context understanding, and fairness. The subsequent sections will delve into the specific components of the proposed framework, discuss potential future improvements, and provide a detailed timetable plan outlining the research process.

2 Related Work

Throughout its development, visual question answering (VQA) has evolved along three main phases: early trials (before 2015), mid progress (2016-2018), and new thoughts (2018-present).

2.1 Early Trials (Before 2015)

In the early stages, researchers began adopting neural networks such as GoogLeNet, CNN, and LSTM for VQA. [6, 7, 9, 14]

2.2 Mid Progress (2016-2018)

As attention-based models achieved considerable success in various fields, researchers aimed to make breakthroughs in VQA. Attention-based models like "Where to Look" [11], "Recurrent Spatial" [15], "Stacked" [13], and "Bottom-Up Top-Down" [2] reached state-of-the-art performance.

2.3 New Thoughts (2018-Present)

Jacob Andreas et al. proposed compositional neural module networks, which improved the interpretability of VQA. [3] Other attempts have also been made in this area. [12]

In 2022, Tanmay Gupta and Aniruddha Kembhavi introduced the Visual Programming framework, shedding new light on the field of VQA. [5] This framework offers new possibilities for exploration.

2.4 Key Challenges

Currently, VQA faces several key challenges, including interpretability and data bias.

Most existing VQA systems lack interpretability, with the exception of two approaches [3, 8]. However, the former's performance is unsatisfactory, and the latter only performs well on limited datasets. Developing a highly interpretable VQA system with state-of-the-art performance is a critical objective.

Another crucial consideration in VQA is the presence of biases in the training data. VQA datasets often exhibit biases towards specific classes, demographics, or visual attributes, leading to biased answers. These biases can stem from the data collection process or societal biases reflected in the training samples.

Biased VQA systems have the potential to perpetuate and amplify societal biases, resulting in unfair and discriminatory answers. Addressing these biases and ensuring fairness in VQA systems is paramount for building inclusive and unbiased AI systems.

G. Kv and A. Mittal [4] and Sasha Sheng et al. [10] have independently identified the bias in the current VQA systems and proposed solutions. They have respectively enhanced methodology and expanded datasets to mitigate biases. Additionally, datasets, baselines, and benchmarks have been proposed to establish standards for metrics. [1]

3 Proposed Approach and Methodology

3.1 Visual Programming

Figure 1: The Framework of VisProg and its capabilities

We could describe the task in natural language and have an AI system generate and execute the corresponding visual program without any training.

From the left part of Figure 1, we can see that the VISPROG mainly consists of 2 parts: Program Generator and Program Interpreter.

With the Program Generator, VISPROG uses the in-context learning ability of GPT-3 to output visual programs for natural language instructions.

Figure 2 shows such a prompt for an image editing task. The programs in the in-context examples are manually written and can typically be constructed without an accompanying image. Each line of a VISPROG program, or a program step, consists of the name of a module, module's input argument names and their values, and an output variable name.

These in-context examples are fed into GPT-3 along with a new natural language instruction. Without observing the image or its content, VISPROG generates a program (bottom of Figure 2) that can be executed on the input image(s) to perform the described task.

As for the Program Interpreter, it translate the generated program into a set of modules' combinations to receive the input image and make correct prediction.

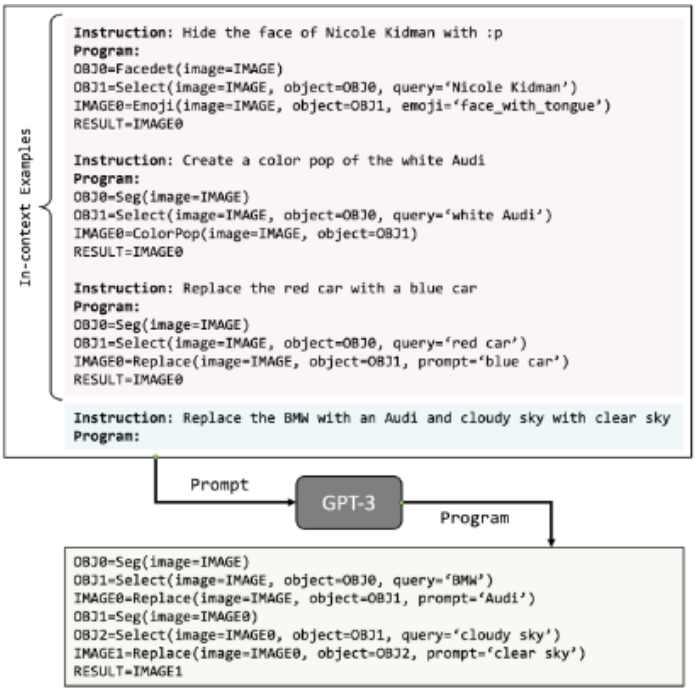

Figure 2: Program Generation in VisProg

We can see the modules currently supported in VISPROG from Figure 3.

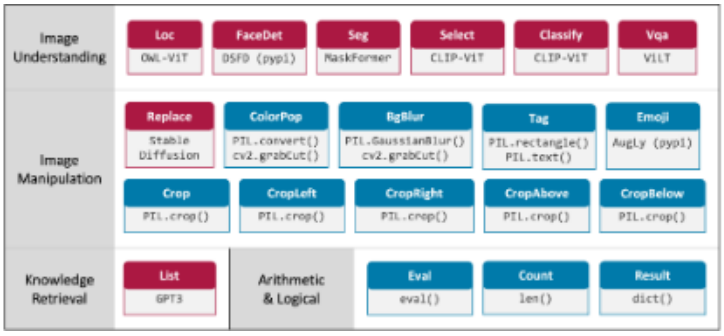

Figure 3: Current available modules in VisProg

3.2 Expected Improvements to Existing Methods

The proposed visual programming framework offers several enhancements to overcome the limitations of existing VQA methods:

Increased Interpretability: Addressing the limited interpretability of VQA models is crucial for building trust and understanding in the generated answers. This enhancement involves the development of mechanisms that provide insights into the internal workings of the VQA model. By visualizing and interpreting

the decision-making process, users can gain a deeper understanding of the rationale behind the answers generated by the system. This improvement promotes transparency and facilitates the identification of potential biases or errors.

Contextual Understanding: Handling complex queries that require contextual understanding is a major challenge in VQA. To address this limitation, the proposed framework incorporates attention mechanisms and contextual processing models. Attention mechanisms enable the model to focus on specific regions of the input image that are most relevant to answering the question. Additionally, contextual processing models enable the system to capture and exploit contextual information effectively. By considering the broader context in which the question is posed, the framework enhances its ability to generate accurate and meaningful answers.

Bias Mitigation: Addressing biases in VQA systems is essential to ensure fairness and to mitigate potential discriminatory outcomes. The framework incorporates techniques to address biases in the training data. This includes careful selection and preprocessing of training data, handling imbalanced class distributions, and incorporating fairness measures during model training. By actively mitigating biases, the proposed framework aims to reduce the risk of biased answers and create more equitable VQA systems.

Flexibility and Adaptability: The visual programming interface of the proposed framework aims to provide users with flexibility and adaptability. This enhancement allows users to customize and extend the VQA framework according to specific requirements. Users can incorporate domain-specific knowledge, add additional functionalities, or integrate specialized modules into the framework. This flexibility empowers users to shape the VQA system to better suit their needs and maximize its performance in various application domains.

Through these improvements, the proposed visual programming framework aims to significantly enhance the accuracy, efficiency, interpretability, context understanding, fairness, and customization capabilities of VQA systems.

4 Goal and Objectives

The proposed visual programming framework for VQA lays the foundation for ongoing research and advancements in the field. Expanding upon this framework, several potential future research directions and improvements can be pursued to further enhance its capabilities and address emerging challenges. These directions include incorporating multi-modal fusion techniques, enhancing robustness, exploring transfer learning approaches, and investigating mechanisms for continuous learning and adaptation.

Incorporating Multi-modal Fusion Techniques: One promising avenue for future research is the integration of multi-modal fusion techniques to leverage additional modalities beyond visual and textual data. VQA can benefit from incorporating modalities such as audio, depth, or haptic signals. Fusion techniques, such as early merging or late fusion, can be explored to effectively combine multiple modalities for a more comprehensive understanding of the visual context. By incorporating additional modalities, the framework can provide richer, more informative answers to visual queries.

Enhancing Robustness: The proposed framework can be further improved to handle variations in visual domains and query types. Robustness to changes

in lighting conditions, viewpoints, or image quality is essential for practical VQA systems. Research efforts can focus on developing methods to ensure the robustness of visual feature extraction, attention mechanisms, and reasoning processes. Adapting the framework to handle diverse query types, such as spatial, temporal, or comparative queries, would allow it to address a broader range of user needs and query variations.

Exploring Transfer Learning Approaches: Transfer learning, which leverages knowledge learned from one task or domain to improve performance in another, holds significant potential for enhancing VQA systems. Future research can explore transfer learning approaches to boost the generalization capabilities of the VQA model. By pretraining the model on large-scale datasets or related tasks, and then fine-tuning it on VQA-specific data, the framework can benefit from shared knowledge and improve its ability to handle unseen or challenging VQA scenarios.

Investigating Mechanisms for Continuous Learning and Adaptation: As new visual data and query patterns emerge over time, VQA systems should be capable of continuous learning and adaptation. Future research can focus on developing mechanisms to enable the framework to evolve and learn from new data. Incremental learning approaches, lifelong learning, or online learning techniques can be explored to ensure the VQA model remains up-to-date and adaptable to evolving contexts. By continuously updating the model with new data, the framework can improve its performance and relevance in real-world VQA applications.

Integration with External Knowledge Sources: Expanding the framework to incorporate external knowledge sources can lead to improved VQA performance. By leveraging external sources, such as knowledge graphs or ontologies, the framework can access a rich repository of contextual information and world knowledge. Integration with external knowledge can enhance the reasoning abilities of the VQA model, enabling it to make more informed and accurate decisions when answering complex queries. Investigating methods to effectively integrate and utilize external knowledge sources will be crucial in advancing VQA systems.

Evaluation and Benchmarking: Establishing comprehensive evaluation metrics and benchmark datasets for VQA systems is essential for comparing and measuring progress. Future research should focus on refining existing evaluation protocols and developing new benchmarks that capture various dimensions of VQA performance, including accuracy, contextual understanding, fairness, and interpretability. The creation of benchmarks covering challenging scenarios or specific domains will provide a standardized framework for evaluating the proposed visual programming framework, as well as other VQA approaches.

By pursuing these future research directions and improvements, the proposed visual programming framework can evolve into a more versatile and powerful tool for managing VQA tasks. Incorporating multi-modal fusion techniques, enhancing robustness, exploring transfer learning approaches, investigating mechanisms for continuous learning and adaptation, integrating external knowledge sources, and establishing comprehensive evaluation benchmarks will push the boundaries of VQA systems and pave the way for their wider adoption in various application domains.

5 Timetable Plan

The following timetable outlines the proposed plan for conducting the research and development of the new visual programming framework for VQA:

December 2023: The initial phase of the research involves thorough literature review and an in-depth exploration of existing VQA methods. This includes studying relevant research papers, reviewing state-of-the-art techniques, and identifying potential gaps and limitations. Additionally, during this month, efforts will be made to collect and preprocess the required datasets for training and evaluation purposes. This involves sourcing diverse datasets that cover a wide range of visual content and associated questions.

January 2024: This month will be dedicated to the development and experimentation phase of the proposed visual programming framework. The focus will be on implementing the different key components, such as visual data integration, image feature extraction, syntax and logic representation, and model integration. A series of experiments will be conducted to evaluate the performance of the framework. Various metrics, including accuracy, efficiency, and interpretability, will be used to assess the efficacy of the proposed approach. The results obtained from these experiments will guide further refinements and iterations of the framework.

February 2024: The primary objective during this month will be to write the first draft of the research paper. The draft will outline the details of the proposed visual programming framework, including its key components and how they address the limitations of existing VQA methods. The findings from the research and experimentation phase will be consolidated, and the framework's potential implications and contributions will be discussed in detail. The draft will also provide an overview of the evaluation results and their significance in the context of the proposed framework.

March 2024: The month of March will involve the mid-assessment stage, during which feedback from experts and evaluators will be sought. The research paper will be refined based on the feedback received, incorporating suggestions and addressing any identified areas of improvement. Additional experiments may be conducted to further validate and extend the findings of the proposed framework. The goal is to ensure that the research paper accurately represents the advancements made in VQA through the visual programming framework.

April 2024: In April, the final assessment of the visual programming framework will be conducted, building upon the refined research paper. The framework's performance and effectiveness will be thoroughly evaluated based on the established evaluation metrics and benchmarks. The assessed results will be incorporated into an enhanced version of the research paper, addressing any remaining feedback or questions raised by the expert evaluators.

May 2024: The final month will be dedicated to making the necessary preparations for the submission of the research paper. This includes thorough proofreading, ensuring consistent formatting, and finalizing the document to meet the required academic standards. The research paper will be submitted, encapsulating the comprehensive exploration of the proposed visual programming framework for VQA, as well as the research findings, insights, and contributions made throughout the research process.

The outlined timetable plan ensures a systematic and comprehensive approach to research, development, evaluation, and dissemination. By adhering

to this plan, the research team can effectively explore the potential of the proposed visual programming framework, contribute to the field of VQA, and offer insights into future advancements and applications.

6 Summary

This research plan has outlined the proposal for a new visual programming framework for managing visual question answering (VQA) tasks. The aim of this framework is to address the challenges faced in accurately and efficiently answering complex visual queries by integrating visual data, programming logic, and question processing capabilities. The significance of this research lies in the potential improvements and future directions it offers to the field of VQA.

The proposed visual programming framework provides a comprehensive approach to VQA by incorporating key components such as visual data integration, image feature extraction, syntax and logic representation, and model integration. By efficiently integrating these components, the framework enables improved accuracy and efficiency in generating answers to visual queries.

The research plan also emphasizes enhancements and future directions for the proposed framework. Increasing interpretability is crucial for building trust and understanding in the generated answers. By developing mechanisms to visualize and interpret the internal workings of the VQA model, the framework aims to provide transparency and facilitate the identification of potential biases or errors.

Contextual understanding is another important aspect that can be enhanced in the proposed framework. The incorporation of attention mechanisms and contextual processing models allows the framework to capture and utilize contextual information effectively, thereby improving its ability to generate accurate and meaningful answers to complex queries.

Additionally, biases in VQA systems need to be mitigated to ensure fairness. Techniques addressing biases in the training data, such as careful selection, preprocessing, and fairness measures during model training, are important considerations for the proposed framework.

Flexibility and adaptability are key attributes of the visual programming interface, which allows users to customize and extend the VQA framework according to specific domain requirements. This flexibility empowers users to shape the framework to better suit their needs and maximize its performance in diverse application domains.

The research plan also provides a clear timeline for the execution of the proposed activities. Over the course of the project, research, experimentation, iterative improvements, and extensive evaluation will be conducted to validate the effectiveness of the visual programming framework for VQA tasks. By adhering to the established timetable, the research team expects to produce a comprehensive research paper by the end of May 2024.

The expected outcomes of this research plan include the development of an advanced visual programming framework for managing VQA tasks, a deeper understanding of the limitations and challenges in existing VQA methods, and insights into future improvements and research directions. The contributions of this research will advance the field of VQA, improve the accuracy and efficiency of VQA systems, enhance interpretability and fairness, and provide a

solid foundation for future developments in managing visual question answering tasks.

In summary, the proposed visual programming framework holds significant potential for managing VQA tasks. As emphasized throughout this research plan, the framework offers improvements over existing methods, particularly in terms of interpretability, contextual understanding, bias mitigation, and customization. By executing the proposed research plan and adhering to the expected timeline, significant advancements can be achieved, contributing to the overall progress of the field and opening new horizons for the design and development of VQA systems.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks, 2017.

[4] A Mittal G Kv. Reducing language biases in visual question answering with visually-grounded question encoder.

[5] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training, 2022.

[6] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network, 2015.

[7] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction, 2015.

[8] Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering, 2018.

[9] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering, 2015.

[10] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[11] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering, 2016.

[12] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering, 2017.

[13] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering, 2016.

[14] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering, 2015.

[15] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016.