Enhance Visual Reasoning in VQA Tasks through Architectural Choices and Training Strategies

Xiumei Xue
University of Gothenburg
gusxuexi@student.gu.se

Abstract

Visual Question Answering (VQA) is a challenging multi-modal task that requires reasoning over both visual and linguistic inputs. This paper investigates the impact of architectural choices and training strategies on VQA performance, focusing on the integration of program-augmented reasoning and online learning. I propose a dynamic feature fusion strategy that integrates program modules and online updates, significantly enhancing model accuracy, particularly for complex reasoning tasks. The results demonstrate the effectiveness of structured reasoning and adaptive learning in improving VQA systems, offering a good foundation for further exploration of interactive learning.

1 Introduction

Visual Question Answering (VQA) is a multi-modal task that bridges computer vision and natural language processing, requiring models to reason over both visual and linguistic inputs. Despite significant advancements in deep learning, VQA remains challenging due to the complexity of aligning visual and linguistic information, especially in dynamic and unseen scenarios.

This paper explores the impact of architectural choices and training strategies on VQA performance, focusing on the integration of program-augmented reasoning and online learning. By examining the influence of varying the maximum program length on three model types: a baseline neural network, a program-augmented model, and a model employing online learning. The results demonstrate significant improvements in validation accuracy, particularly when program modules are integrated. This study provides insights into the interplay between architectural design and training strategies, highlighting the potential of structured reasoning and adaptive learning for enhancing VQA systems. Future directions include exploring more advanced architectures and integrating external knowledge to further improve model robustness and generalization.

2 Related Work & Datasets

2.1 Visual Question Answering

Visual Question Answering (VQA) has emerged as a critical area of research in artificial intelligence, requiring models to integrate visual and linguistic information for complex reasoning tasks. Since its inception in the mid-2010s, VQA has evolved from a niche topic to a cornerstone of AI applications, with tasks ranging from simple object recognition to advanced spatial reasoning and common-sense understanding.

Despite its potential in areas such as assistive technologies, education, and medical image analysis, VQA faces significant challenges. These include the need for contextual understanding, dataset biases, and the difficulty of aligning multi-modal information. Recent work has attempted to address these issues through advanced architectures and training strategies, yet gaps remain in handling complex reasoning tasks and dynamic scenarios. In this paper, we build on these efforts by exploring the integration of program-augmented reasoning and online learning to enhance VQA performance.

2.2 Interactive Learning

From our readings and discussions so far, one of the key observations I have made is that current neural network models struggle with tasks requiring spatial reasoning or counting. They often rely on pattern recognition rather than true spatial un-

derstanding. Given this, I am somewhat skeptical of the research path based on the inherent, restrictive approach, and am more drawn to exploring interactive learning.

Interactive learning is a paradigm in artificial intelligence where agents learn to perform tasks through interactions with a teacher or environment. This approach is particularly relevant in scenarios where the agent must adapt to dynamic or unforeseen changes in its operating context.

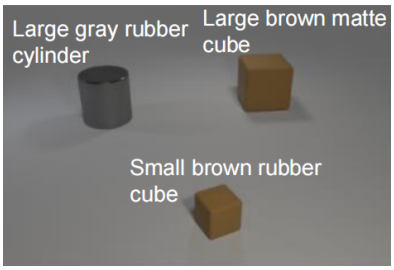
In my project, the data set is static, and I want to explore whether adding human feedback such as corrections can improve accuracy. Feedback could be incremental rather than a direct answer. However, I encountered practical implementation challenges, such as the mismatch between batch-encoded inputs and item-specific human corrections in natural language, which requires significant manual efforts. Given the time constraints, the focus of the project has a little conversion. I mainly compare the results of different architectural choices and training strategies.

2.3 Available Models

The development of VQA systems has been supported by advancements in both vision and language processing models.

On the vision side, convolutional neural networks (CNNs) like VGG (Simonyan and Zisserman, 2015) and ResNet-50 (He et al., 2015) have played pivotal roles. VGG is known for its simplicity and uniform architecture, using sequential convolutional layers to extract hierarchical image features. ResNet-50, on the other hand, introduced residual connections, enabling the training of much deeper networks by mitigating vanishing gradient issues. These models excel in visual feature extraction, making them widely used in VQA systems to process image data.

For language understanding, recurrent neural networks (RNNs) such as LSTM (Hochreiter and Schmidhuber, 1997) were traditionally used to handle sequential data. LSTMs effectively capture long-range dependencies, making them suitable for encoding questions in VQA tasks. However, more recently, transformers like BERT (Devlin et al., 2019) have revolutionized natural language processing. BERT leverages self-attention mechanisms and large-scale pre-training to achieve state-of-the-art performance in various NLP tasks. Its contextualized embeddings significantly enhance a model's ability to understand complex and ambiguous questions.



Q: There is a rubber cube in front of the big cylinder in front of the big brown matte thing; what is its size?
A: small
Q-type: query_size
Size: 14

Q: What color is the object that is on the left side of the small rubber thing?
A: gray
Q-type: query_color
Size: 7

Figure 1: Example image and QA pairs in CLEVR.

These foundational models have been adapted and extended in various ways to address the unique challenges of VQA tasks, such as integrating visual and textual information or improving reasoning capabilities.

2.4 CLEVR Dataset

To promote the development of VQA research, researchers have developed various datasets, among which the CLEVR (*Compositional Language and Elementary Visual Reasoning*) dataset is a significant contribution.

The CLEVR dataset is a comprehensive dataset designed to test the visual understanding and reasoning capabilities of AI systems (Johnson et al., 2016). It is specifically designed to evaluate the model's capabilities in attribute recognition, counting, comparison, and spatial relationship reasoning by providing a series of synthetic 3D rendered images and corresponding complex questions (See example in Figure 1). CLEVR is unique in that its questions require multi-step reasoning, making it an ideal tool for evaluating the deep understanding and reasoning capabilities of VQA systems.

The CLEVR dataset was created to address some of the limitations of earlier VQA datasets, such as data bias and insufficient reasoning depth. By using comprehensive images and questions, CLEVR ensures data diversity and balance, while also providing groud-truth annotations (scene graphs)

Split	Images	Questions	Unique questions	Overlap with train
Total	100,000	999,968	853,554	-
Train	70,000	699,989	608,607	-
Val	15,000	149,991	140,448	17,338
Test	15,000	149,988	140,352	17,335

Table 1: Statistics for CLEVR; the majority of questions are unique and few questions from the val and test sets appear in the training set.

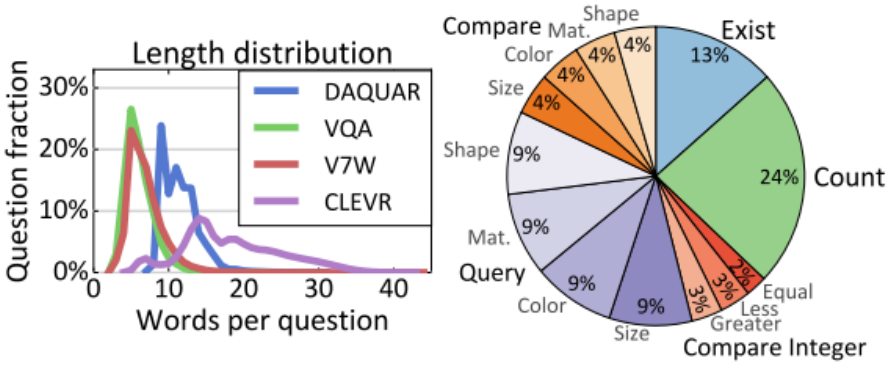


Figure 2: Left - CLEVR questions are generally much longer. Right - Distribution of question types in CLEVR.

and functional programs, allowing researchers to deeply analyze the model's reasoning process. Since its release, CLEVR has become one of the important benchmarks for evaluating the performance of VQA models, driving the rapid development of this field.

3 Methodology

3.1 Baseline Model

The initial baseline model follows the lecture, which utilizes an LSTM model to process questions and VGG-16 to extract image features.

Since the output effect was not ideal, I upgraded the model structure, replacing VGG-16 with ResNet-50 and the LSTM encoder with BERT. ResNet-50 improves upon VGG by introducing residual connections, enabling deeper networks with better generalization capabilities. BERT, as a transformer-based model, replaces the sequential processing of LSTM with parallelized self-attention mechanisms, allowing for more nuanced understanding of contextual relationships in questions. These upgrades aim to enhance both the visual and textual understanding components of the model.

3.2 Program-Augmented Model

Program is a special attribute in the CLEVR dataset and it is somewhat similar to the current popular concept of *Chain-of-Thinking* (CoT). For example, it can parse a single question into multiple steps such as 'query_size' and 'count.' As the length of program grows, the complexity of questions increases. By combining neural networks with program reasoning modules, the model could be enhanced with intermediate program representations.

In terms of combination, the preliminary method is a simple concatenation. To make this integration more flexible, I adopt a dynamic fusion policy, where the importance of each component (images, questions, and programs) is adjustable during the training or learning process. This integration leverages the strengths of each component: ResNet-50 for detailed image understanding, BERT for robust question interpretation, and program modules for structured reasoning.

3.3 Online Learning

I extend the program-augmented model with online updates during inference to adapt dynamically. The online learning process allows the model to iteratively refine its parameters based on incoming data or feedback, enabling improved generalization to new or unseen scenarios.

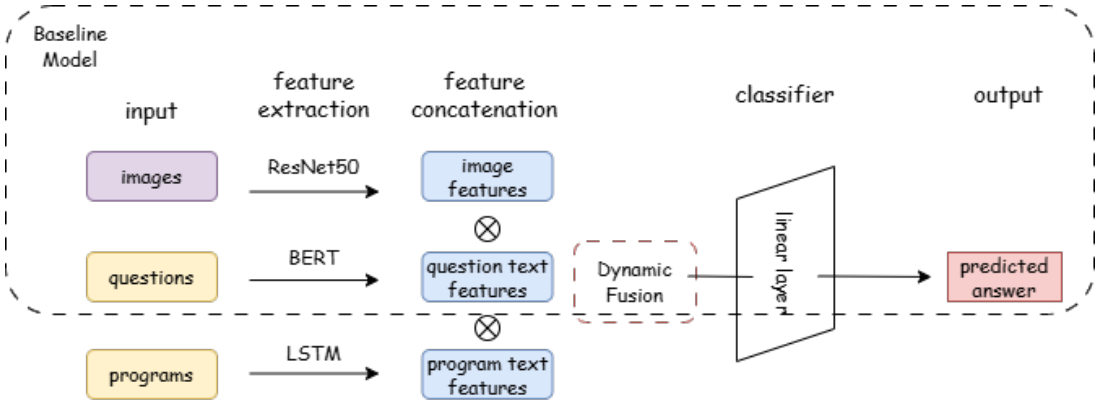


Figure 3: Model Architecture.

This extension further builds upon the capabilities of the underlying components. ResNet-50 and BERT provide strong initial feature extraction, while the program reasoning module supports intermediate step-by-step reasoning. Online learning introduces an adaptive layer to the system, enabling dynamic updates that align with the real-time context or task-specific requirements.

4 Results and Discussion

I started my experiment[1] with a simple model, and the output was not ideal–the accuracy on validation set was only around *3%*.

After the model structure optimization, a noticeable improvement in accuracy[2] was observed. The best-performing results are summarized in Table 2. From the analysis, the following key insights can be drawn:

1) The performance of model is relevant to the complexity of questions. By controlling the program length, it becomes evident that performance degrades as the questions become more complex. This suggests that the model struggles with intricate queries, highlighting a limitation in handling higher levels of abstraction or detail.

2) Programs serve as valuable additional information, particularly for relatively complex questions. Though both training and validation accuracies decrease when *max_program_length* is set as 5, the program-enhanced model still demonstrates a significant improvement over the baseline. This indicates that incorporating programmatic information can enhance the model's ability to process and reason about more challenging tasks.

3) The generalization capability of model is limited. The results reveal a notable gap between training and validation accuracy, which points to potential over-fitting. This discrepancy underscores the model's limited ability to generalize to unseen data. Moreover, the gap highlights the complex nature of VQA tasks and implies the potential value of incorporating human feedback to improve the model's adaptability and robustness in real-world scenarios.

Additionally, while the results of online learning appear promising, they exhibit a degree of randomness. This suggests that the online learning process may lack stability or consistency, warranting further investigation into its reliability and effectiveness.

5 Conclusion

This preliminary work highlights the impact of *max_program_length* and architectural enhancements on neural network performance. By integrating program reasoning and online learning, models achieve superior accuracy in VQA tasks.

In terms of model architecture, recent progress in deep learning, particularly in areas such as transformer architectures and large-scale pre-training, have pushed the improvements of VQA performance. Nevertheless, there remain substantial opportunities for improvement in areas such as generalization to novel scenarios, and the integration of external knowledge. These limitations highlight the need for continued innovation in model design and training methodologies.

What the work highlights is that, though in most real VQA datasets, the questions lack programming annotations, we can enhance the input with similar ideas. (Shah et al., 2019) propose VQA-

[1]You can find the project code here: https://github.com/xxmorwellr/AICS-project-2024

[2]Due to the absence of the *program* attribute in the test set, our analysis focuses on the training and validation sets

Model	max_program_length	Train Accuracy (%)	Val Accuracy (%)
Baseline Model	3	89.10	48.79
Enhanced With Program	3	95.77	47.57
Online Learning	3	/	62.50
Baseline Model	5	26.04	18.41
Enhanced With Program	5	60.99	46.67
Online Learning	5	/	68.75

Table 2: Performance comparison of different models and subsets

Rephrasings, a new VQA dataset and evaluation protocol to measure robustness of VQA models to linguistic variations, and a new metric called *cycle-consistency*. It implies the potential of rephrasing questions, especially considering the complexity and ambiguity of natural languages. Özdemir et al.(2024) incorporate image captioning as an intermediary process within the VQA pipeline, which is also inspiring to enhance the questions.

6 Future Work

Future work will focus on exploring more advanced architectures and hyperparameter configurations. Given the inherent limitations of CLEVR dataset, to validate the model on hybrid datasets that combine synthetic and real-world data, also help assess the model's ability to generalize beyond synthetic environments.

Additionally, I still aim to investigate the potential of interactive learning, with a particular emphasis on improving the integration of real-world knowledge into VQA systems. This direction holds promise for enhancing the adaptability and practical applicability of such models in real-world scenarios.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.

John E. Laird, Kevin Gluck, John Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario Salvucci, Matthias Scheutz, Andrea Thomaz, Greg Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive task learning. *IEEE Intelligent Systems*, 32(4):6–21.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. *CoRR*, abs/1902.05660.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Övgü Özdemir and Erdem Akagündüz. 2024. Enhancing visual question answering through question-driven image captions as prompts.