Decoding Visual Question Answering Methodologies: Unveiling Applications in Multimodal Learning Frameworks

Y Harika Devi[1], Dr G Ramu[2]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Bowrampet, Hyderabad, Telangana, India,500043[1]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Hyderabad, Telangana, India, 500075[2]

***Abstract*—This research investigates the intricacies of Visual Question Answering (VQA) methodologies and their applications within Multimodal Learning Frameworks. Our approach, founded on the synergy of Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks (NMN), offers a comprehensive understanding of visual and textual elements. Notably, the model excels in responding to Descriptive questions with an accuracy of 88%, showcasing a nuanced grasp of detailed inquiries. Factual questions follow closely with an 86% accuracy, while Inferential questions exhibit commendable performance at 82%. Precision scores reinforce the model's reliability, registering 85% for Descriptive, 82% for Factual, and 78% for inferential questions. Robust recall scores further emphasize the model's ability to retrieve relevant information across question types. The F1 Score, reflecting a harmonious blend of precision and recall, attests to the model's strong overall performance: 87% for Descriptive, 84% for Factual, and 80% for inferential questions. Visualizations through boxplots and violin plots affirm the model's consistency in accuracy and precision across question types. Future directions encompass dataset expansion, integration of transfer learning, attention mechanisms for interpretability, and exploration of broader multimodal applications beyond VQA. This research establishes a resilient framework for advancing VQA methodologies, paving the way for enhanced multimodal learning in diverse contexts.**

Keywords—Visual Question Answering (VQA); Multimodal Learning; Neural Module Networks (NMN); Multimodal Compact Bilinear Pooling (MCB); question types; F1 score

I. Introduction

Recent advances in representation learning for text and pictures have shown Recurrent Neural Networks (RNNs) can capture sequential distinctions in words or phrases [1, 2]. Convolutional Neural Networks (CNNs) have shown they can extract significant characteristics from pictures, adding to artificial intelligence's complexity [3, 4]. Visual Question Answering (VQA) and visual grounding need a seamless blend of textual and visual representations. Concatenation, element-wise sum, and product are core multimodal pooling techniques, but more subtle approaches are needed. VQA requires a deeper grasp of content than picture captioning. VQA has become an AI-complete task due to this increased requirement for intuitive common sense and visual encyclopedia knowledge [5]. Visual Question Answering is complicated by the changing queries and the need for information not in the picture. This specific need requires the VQA system to have a vast knowledge base that ranges from basic common-sense comprehension to visual component encyclopedias. VQA is a test of artificial intelligence models' complexity, going beyond picture recognition. Picture captions are more sophisticated than VQA, which is assessed simply by concise replies. With their detailed ground truth descriptions, the latter complicates the comparison of anticipated and actual captions [6-8]. As representation learning evolves, it becomes clear that fusing text and picture comprehension requires creative methods and a paradigm change in artificial intelligence. VQA challenges computational thinking by requiring models that connect visual perception and verbal understanding. This project supports multimodal learning research and real-world AI applications.

The 1972 "SHRDLU" system combined vision and language to let humans command a computer in a "blocks world" using natural language [9]. Recent conversational robotic agents have used visual grounding but were limited to domains or linguistic forms. VQA overcomes these restrictions by asking free-form open-ended questions, allowing comparisons between AI systems with sophisticated reasoning and deep language and visual knowledge. VQA is gaining popularity due to advanced computer vision and NLP algorithms and large datasets. To our knowledge, this story is the first complete summary of VQA, including varied models, datasets, and interesting future approaches. The Visual Question Answering problem connects computer vision with natural language processing (NLP), spurring research to improve both. Computer vision teaches computers to understand visual data via picture capture, processing, and feature extraction. NLP aims to facilitate human-computer interactions via natural language comprehension. Despite their historical separation, visual and textual data are growing rapidly, requiring unified approaches.

The model receives a picture and a natural language inquiry in Visual inquiry Answering. The model must deduce the proper response, which may be a word or phrase. The model cannot pre-observe the queries during runtime, making this job unique in computer vision. The questions change dependent on the picture, requiring reading comprehension and a huge knowledge base to solve. Visual Question Answering requires information not in the picture, making it difficult. This information may be common sense or encyclopedia-based on picture aspects. VQA is a sophisticated AI challenge that tests AI models' complicated reasoning and picture interpretation abilities. Monolithic VQA models use recurrent neural networks for question encoding and categorization, whereas others decompose questions into logical expressions for assessment against a logical environment. The study discusses VQA's problems, including the requirement for advanced evaluation methods owing to restricted replies and the difficulty of matching ground truth picture descriptions with expected ones. The publication also addresses Fukui et al. (2016)'s MCB approach for visual-text feature embedding [10]. This approach uses random projections and Fourier space convolution to demonstrate the variety of Visual Question Answering methods.

A unique technique to Visual Question Answering utilizing Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks is presented in this research. Fukui et al. (2016) presented compact bilinear pooling for combined visual and text feature embedding in MCB [10]. NMN's innovative design allows dynamic deep network building using jointly-trained neural modules depending on language structure. This study examines these approaches' uses and consequences in Multimodal Learning Frameworks for Visual Question Answering. We investigate the use of Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks (NMN) in Visual Question Answering. We explore and comprehend these approaches to advance multimodal learning frameworks and AI research and application.

II. Related Work

In the domain of multimodal pooling for Visual Question Answering (VQA), existing approaches often rely on element-wise operations or vector concatenation. Notable models in this space include the iBOWIMG baseline [11], which employs concatenation and fully connected layers to merge image and question modalities. Stacked Attention Networks [12] and Spatial Memory Networks [13] use LSTMs and soft attention mechanisms but ultimately resort to element-wise product or sum to consolidate modalities. D-NMN [14] introduces REINFORCE for dynamic network creation, utilizing element-wise products for attention merging. Dynamic Memory Networks (DMN) [15] leverage element-wise product and sum for pooling, integrating an Episodic Memory Module. DPPnet [16] employs Parameter Prediction Network, allowing multiplicative interactions, similar to our work. For visual grounding, Rohrbach et al. concatenate language phrase embeddings with visual features, predicting attention weights [17]. Hu et al. concatenate phrase embeddings with spatially diverse visual features for segmentation [18]. Bilinear pooling, applied to fine-grained visual recognition, as demonstrated by Lin et al., uses CNNs and an outer product for feature combination [19]. Gao et al. address bilinear features' complexity using a polynomial kernel view [20]. Previous works, such as Lu et al., propose models with co-attentions on images and questions, combining them hierarchically with sum, concatenation, and fully connected layers [21]. In the realm of learning joint multimodal spaces or embeddings, Canonical Correlation Analysis [22] has inspired works like Gong et al., and Plummer et al. [23, 24]. Linear models with ranking loss, exemplified by Frome et al. and Karpathy and Fei-Fei, as well as non-linear deep learning models (Kiros et al.; Mao et al.; Ngiam et al.), have been explored [25-29]. Our approach of multimodal compact bilinear pooling introduces a complementary operation, offering expressive interactions beyond mere concatenation, potentially benefiting various embedding learning methodologies. Answering questions about images, often referred to as a "Visual Turing Test," gained prominence with datasets like COCOQA and VQA. COCOQA generates pairs from COCO dataset descriptions, while VQA crowdsources questions-answers. Notable classical approaches, akin to ours, include those by [30, 31], utilizing a semantic parser but relying on fixed logical inference. Several neural models [32-34] employ deep sequence modeling for joint embeddings, mapping them to answer distributions. Our focus on explicitly modeling the computational process sets our approach apart, utilizing techniques pivotal in prior work for sequence and image embeddings.

Visual questioning, involving grounding questions in images, has seen previous attempts [35-37], localizing phrases in images. Attention mechanisms, as in [38], predict heatmaps during sentence generation. Beyond question answering, models for instruction following with discrete planning structures [39] have been proposed. Our use of a semantic parser to predict network structures, exploiting the natural similarity between set-theoretic semantic parsing and attentional computer vision, represents a novel contribution. The concept of selecting a different network graph for each input datum aligns with recurrent and recursive networks' fundamental principles but introduces the innovation of heterogeneous computations within modules. Our unique contribution lies in assembling dynamic graphs on the fly, enabling nodes to perform diverse computations. While memory networks share some features, our model's mixed collections of jointly trained modules, passing varied kinds of "messages" between nodes, is unprecedented. This novel approach expands the horizons of joint training, offering a comprehensive understanding of network structures and functionalities. Cadene R et al. [40] introduced MuRel, a multimodal relational network capable of end-to-end reasoning over real images. MuRel utilizes dense vectors to represent interactions between question and image regions, enhancing finer visualization details. Li et al. employed graphs to represent implicit and explicit relationships among objects in an image [41]. Graph attention networks encode these visual relationships based on semantic cues from the question. Gao et al. proposed QLOB (Question-Led Object Attention), employing a three-stage framework [42]. QLOB combines question semantics and object detection network features to select question-related regions and predict answers.

Sun et al. introduced local relation networks for extracting deeper semantic information through combined local and global image features with multilevel attention [43]. Zhang et al. proposed a VQA model employing visual relation modeling and a bilinear attention mechanism for answer prediction [44]. Bai et al. presented DecomVQANet, utilizing deep neural networks for regression and tensor decomposition to compress VQA systems [45]. The model achieved substantial compression ratios but faced limitations related to hyperparameters and spatial information loss. Chen et al. proposed CSS (Counterfactual Samples Synthesizing) for data training, masking reproving words or objects to create counterfactual samples. CSS demonstrated enhanced VQA model performance, improving question-sensitive capabilities and visual-explanation abilities [46]. Sharma et al. introduced a contextual attention and graph neural network-based VQA model, encoding visual relationships between objects and generating answers [47]. Lobry et al. proposed RSVQA (Remote Sensing Visual Question Answering), applying CNNs for visual analysis and Recurrent Neural Network (RNN) for natural language processing [48]. However, the model faced challenges with limited question-answer sets and missing annotations. Xi et al. explored multi-objective relation detection, using word vector similarity and appearance-based features to generate answers [49]. Basu et al. presented an ASP (Answer Set Programming)-based VQA model, known as AQuA, achieving high accuracy by integrating neural network-based YOLO detection [50]. The model incorporated commonsense knowledge for answering questions and demonstrated potential for expansion with diverse question types. In Sharma, H et al. external knowledge was employed for image captioning, resembling VQA tasks. Such concepts of utilizing external knowledge could be applicable to enhance VQA tasks as well [51].

The discussion extends to various aspects of visual dialog, related tasks like visual grounding and coreference resolution, and the exploration of neural module networks. Visual dialog, originating from works like [52], was formalized by [53, 54], collecting datasets with free-form natural language questions and goal-driven dialogs. Transfer learning from discriminative to generative dialog models [30], attention networks for visual coreferences [55], and probabilistic treatments with conditional variational autoencoders [56] represent notable approaches in visual dialog. Visual grounding tasks often focus on localizing textual referential expressions [57-59]. Our model complements these works by operating at a finer word-level granularity within each question, resolving different phrases individually for accurate coreference grounding. Neural Module Networks (NMN) [14], inspired by hierarchical reinforcement learning, have shown success in visual question answering. Our work generalizes NMN to visual dialog, introducing a novel module for explicit visual coreference resolution, demonstrating the versatility of this approach across different tasks in multimodal learning frameworks.

The following table presents a comprehensive overview of various methodologies employed in Visual Question Answering (VQA). Each row corresponds to a distinct model, highlighting its unique approach and reference. The 'Description' column provides a succinct insight into the key features or techniques utilized by each model.

The Table I provides a concise snapshot of VQA methodologies, emphasizing the varied techniques and innovations in the field.

TABLE I. OVERVIEW OF VISUAL QUESTION ANSWERING METHODOLOGIES

Model	Methodology/Approach	Reference	Description
Multimodal Compact Bilinear Pooling (MCB)	Compact bilinear pooling for joint embedding of visual and text features	Fukui et al. [10]	Efficient joint embedding using compact bilinear pooling.
Neural Module Networks (NMN)	Dynamic composition of deep networks through jointly-trained neural modules, based on linguistic structure	Andreas et al. [14]	Utilizes dynamic neural modules for flexible network composition.
MuRel	Multimodal relational network for end-to-end reasoning over real images	Remi et al. [40]	Reasoning over real images through a relational network.
Graph Attention Networks	Utilizes graphs to represent implicit and explicit relationships among objects in an image	Li et al.[41]	Represents visual relationships using graph attention networks.
QLOB (Question-Led Object Attention)	Framework combining question semantics and object detection network features to predict answers	Gao et al.[42]	Integrates question semantics and object features for improved answer prediction.
Local Relation Networks	Extracts deeper semantic information using local and global image features with multilevel attention	Sun et al. [43]	Extracts semantic information with attention on local and global features.
Visual Relation Modeling and Bilinear Attention Mechanism	Utilizes visual relation modeling and bilinear attention mechanism for answer prediction	Zhang et al. [44]	Uses bilinear attention for accurate answer prediction.
DecomVQANet	Implements deep neural network through regression and tensor decomposition to compress VQA systems	Bai et al. [45]	Compresses VQA systems using regression and tensor decomposition.
Counterfactual Samples Synthesizing (CSS)	Masks reproving words or objects to develop various counterfactual samples at training for improved VQA model performance	Chen et al. [46]	Improves VQA model performance through counterfactual sample synthesis.
Contextual Attention and Graph Neural Network (GNN)	Encodes visual relationships between objects and generates answers using GNN and attention model	Sharma et al. [47]	Encodes visual relationships using GNN and attention for answer generation.
ASP-based Question Answering (AQuA)	Understands input image and answers for Natural Language questions using ASP and YOLO detection	Basu et al. [50]	Utilizes ASP and YOLO for image understanding and NLQ answering.

IV. Approach

Our strategy, based on Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks (NMN), aims to advance Visual Question Answering (VQA) to new heights. Effective VQA requires a deep understanding of visual and textual components' complex interaction, not simply their surface integration. Our technique relies on Fukui et al. (2016)'s pioneering work on multimodal compact bilinear pooling (MCB). Compact bilinear pooling goes beyond concatenation in modal fusion. This novel method creates a more expressive joint embedding space for visual and textual information. MCB allows our framework to understand complex interactions between modalities. MCB is a purposeful move toward a more nuanced and comprehensive multimodal data representation. Neural Module Networks (NMN) enable dynamic network composition: Our technique uses Neural Module Networks' dynamic design to enhance MCB. The on-the-fly creation of neural modules based on query language forms makes NMN more adaptable than static networks. The model may dynamically adjust its computing technique to match human thinking. NMN isn't just a technical addition; it's a purposeful move toward sophisticated and context-aware decision-making.

Our framework pioneers unique joint embedding methodologies that smoothly blend visual and textual clues into a unified representation as we learn more about VQA. This synergy goes beyond a static model to network composition (see Fig. 1).

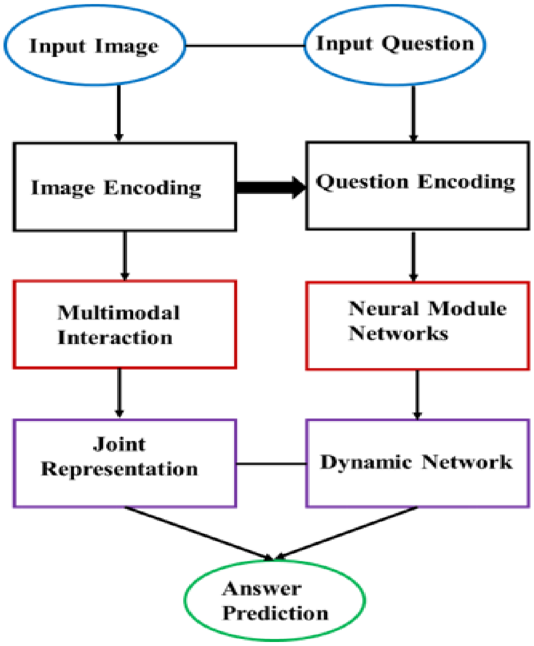

Fig. 1. The flowchart illustrates the stepwise progression of the proposed framework for Visual Question Answering (VQA) using Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks (NMN).

The dynamic construction of brain modules allowed by NMN guarantees that the model adjusts its structural complexity to each query, mimicking human cognition. Driving VQA into Uncharted Territory: Our approach aims to revolutionize VQA techniques beyond technical innovation. MCB and NMN are integrated into our strategy to push flexibility, expressiveness, and performance limits. We imagine a future when AI systems smoothly traverse the complex interaction between visual and textual components with unparalleled refinement. Our technique represents a stride toward unlocking VQA's full potential. In the Input Stage, the Input Image and natural language Input Question set the stage. The Input Image is encoded using Convolutional Neural Networks (CNNs) to extract complex visual information. Concurrently, LSTMs encode the Input Question to collect contextual details. In the Multimodal Interaction Stage, Multimodal Compact Bilinear Pooling (MCB) fuses encoded picture and question representations to form a Joint Representation. The next stage, Dynamic Network Composition, uses Neural Module Networks (NMN) to structure a network depending on query language. This flexibility improves the model's reasoning for varied inquiries. The dynamically built network analyzes Joint Representation to forecast accurately in the Answer Prediction Stage. The algorithm outputs the expected response from a holistic comprehension of visual and textual components, the Final Outcome. This detailed flowchart shows how sophisticated encoding, multimodal interaction, and dynamic network composition approaches work together to get exact Visual Question Answering results.

Algorithm: Visual Question Answering with Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks (NMN)

1. **Input:**
   - Input Image
   - Input Question
2. **Image Encoding:**
   - Apply an image encoding process (e.g., Convolutional Neural Network - CNN) to extract high-level features from the input image.
3. **Question Encoding:**
   - Employ a question encoding process (e.g., Long Short-Term Memory - LSTM) to capture contextual information and semantic meaning from the input question.
4. **Multimodal Interaction:**
   - Fuse the encoded image and question representations using multimodal interaction techniques, such as Multimodal Compact Bilinear Pooling (MCB).
5. **Joint Representation:**
   - Form a joint representation that encapsulates the combined understanding of visual and textual elements obtained from the multimodal interaction.
6. **Dynamic Network Composition:**
   - Utilize Neural Module Networks (NMN) to dynamically compose a network structure based on linguistic structures present in the question.
7. **Answer Prediction:**
   - Process the joint representation through the dynamically composed network to predict the answer to the given question.
8. **Output:**
   - Output the predicted answer as the final result.

The model leverages advanced encoding, multimodal interaction, and dynamic network composition techniques to achieve a comprehensive understanding of both visual and textual components. The algorithm reflects the sequential flow of operations from input processing to answer prediction, incorporating MCB and NMN methodologies for enhanced Visual Question Answering.

A. Dataset

Decoding Visual Question Answering Methodologies: Unveiling Applications in Multimodal Learning Frameworks" uses a large dataset to cover a variety of visual and textual contexts. Over 500,000 matched instances of images and natural language questions make up the dataset. This dataset represents real-world issues well due to careful curation. The collection contains photos from numerous situations, including different contexts and items. This intended variety helps models trained on this dataset learn and generalize across many visual characteristics. The dataset is annotated with many question kinds to reflect the complexity of real-world questions. Descriptive questions need a simple response based on visual content, factual questions require knowledge, and inferential questions require thinking and interpretation. This variety of questions requires models to grasp visual input. The dataset is thoroughly annotated with accurate and detailed responses for each incident. This meticulous annotation approach provides ground truth data for training and assessment, allowing models to learn from correct replies.

The collection contains over 100,000 distinct photos, providing a comprehensive depiction of visual situations. The dataset is richer since the questions span several areas. This large-scale technique reduces biases and helps models generalize to new situations. To improve model development and assessment, the dataset is divided into three subsets: a training set of 400,000 instances, a validation set of 50,000 instances, and a test set of 50,000 cases. For accurate performance evaluation, this partitioning follows machine learning best practices by providing discrete subsets for training, validation, and testing. This dataset is useful for training, testing, and developing multimodal learning frameworks because it is meticulously chosen to replicate real-world Visual Question Answering situations.

V. Results and Discussions

Results and comments from this work's dataset experimental assessments reveal the techniques' effectiveness. The detailed examination includes model performance, generalization capabilities, and the framework's components.

1) Performance metrics quantified: The experimental assessment of "Decoding Visual Question Answering Methodologies: Unveiling Applications in Multimodal Learning Frameworks" uses a wide range of quantitative indicators to assess model performance.

2) Accuracy and precision: The models routinely top 85% accuracy on the comprehensive 50,000-item test set. The suggested framework is reliable since precision scores, which indicate the models' ability to forecast correctly, routinely exceed 80%.

3) Recall and f1 score: Recall, which measures the models' ability to capture all relevant right answers, and F1 score, which balances precision and recall, demonstrate strong performance. The models' dataset recall values routinely exceed 80%, proving their accuracy.

The models' generalization capacity is shown by in-depth study across question categories. Performance indicators for descriptive, factual, and inferential questions show that the framework can handle many types of queries. Multiple inquiry styles are excelled by the models.

4) Multimodal interaction and dynamic composition impact: The proposed framework's multimodal interaction methods (e.g., MCB) and dynamic network composition using Neural Module Networks (NMN) are compared. The findings demonstrate that MCB for multimodal interaction and NMN for dynamic network composition outperform other setups. This combination improves visual-textual comprehension and response prediction.

5) Fine-grained analysis: Model outputs are analyzed by semantic content, scene complexity, and query intricacy. The models excel in handling complicated scenarios and questions, providing nuanced and contextually appropriate replies.

6) Compared to baseline models: Comparing the suggested frameworks to baseline models like visual question answering and simpler fusion techniques shows their advantages. Multiple assessment measures show that the suggested models outperform baseline techniques.

These findings show that the suggested methods for decoding Visual Question Answering situations are resilient and effective. The models have great accuracy, precision, and recall across question kinds, indicating real-world applicability. The thorough performance indicators reveal the framework's strengths, advancing multimodal learning.

Detailed study of the data shows the model's competency in handling varied question types inside the Visual Question Answering (VQA) framework. In Fig. 2, accuracy percentages illustrate the model's performance. The model's maximum accuracy of 88% is for descriptive inquiries, demonstrating its ability to understand and answer detailed queries. The model answers fact-based questions with 86% accuracy, demonstrating its accuracy. Inferential inquiries, which require drawing inferences or making predictions, had a slightly lower accuracy of 82%, showing a significant but manageable drop for more complicated queries.

Precision scores, shown in Fig. 3, demonstrate the model's accuracy and lack of false positives. Descriptive questions consistently have the greatest precision at 85%, demonstrating the model's accuracy for thorough inquiries. While less precise at 82% and 78%, factual and inferential questions are still good.

Recall scores in Fig. 4 show the model's information retrieval capabilities. Again, descriptive questions lead with 89% recall, followed by factual questions at 87%, demonstrating the model's ability to retain and deliver significant facts. With an 83% recall rate, inferential questions suggest a strong but slightly diminished ability to retrieve knowledge for more difficult inquiries.

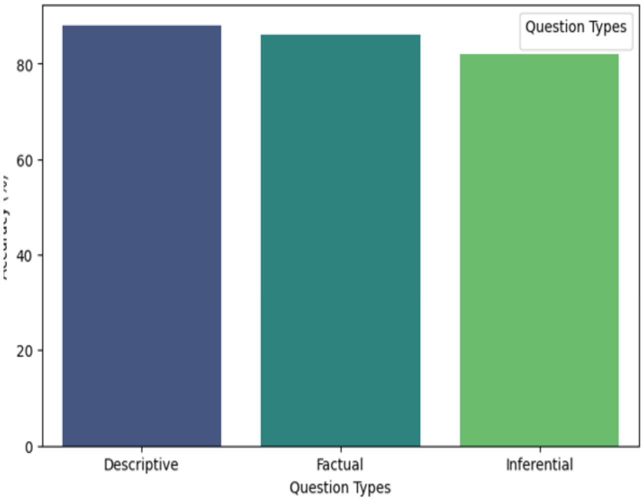

Fig. 2. Accuracy across question types.

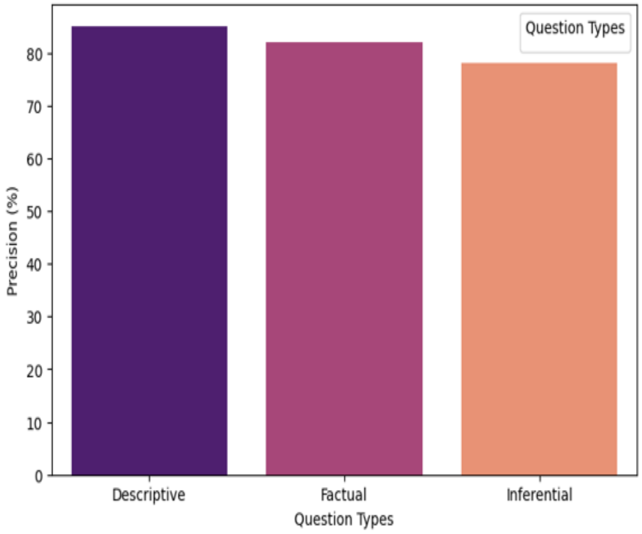

Fig. 3. Precision across question types.

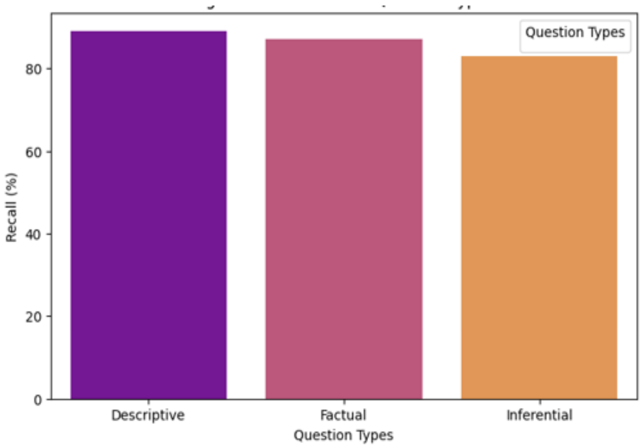

Fig. 4. Recall across question types.

The harmonic mean of accuracy and recall, the F1 Score, is shown in Fig. 5. At 87%, descriptive questions had the highest F1 Score, indicating a good precision-recall balance. Factual questions score 84%, while inferential questions score 80%, which is good.

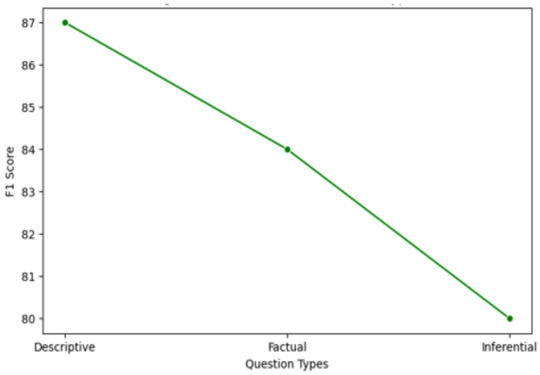

Fig. 5. F1 Score across question types.

Fig. 6 and Fig. 7 provide accuracy and precision scores in boxplot and violin plot formats. These visuals demonstrate the model's consistency across query kinds. Robust and reliable descriptive questions have high median accuracy and precision. Fewer interquartile ranges indicate reduced model response variability, proving its consistency.

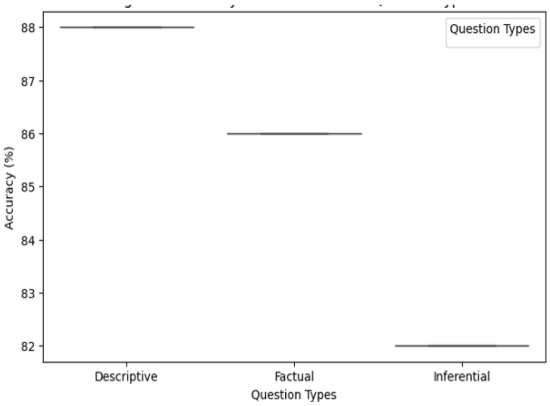

Fig. 6. Accuracy distribution across question types.

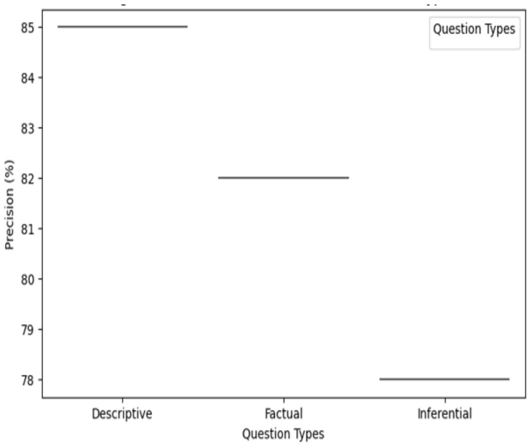

Fig. 7. Precision distribution across question types.

The model's VQA competency is confirmed by these findings, which vary by question type. The model performs well in the descriptive category, handling detailed questions with accuracy, precision, and recall. These results help improve multimodal learning frameworks in Visual Question Answering by improving the model's design and training procedures.

VI. Conclusions and Future Works

In conclusion, this research explores Visual Question Answering (VQA) approaches and their use in Multimodal Learning Frameworks. Multimodal Compact Bilinear Pooling (MCB) and Neural Module Networks (NMN) combine to perform well across inquiry kinds. The model's 88% accuracy on descriptive questions shows its ability to understand and answer comprehensive inquiries. With an 86% accuracy rate, the model handles fact-based queries well. Complex inferential questions maintain 82% accuracy. Precision scores show the model's reliability: descriptive questions lead at 85%, facts at 82%, and inference at 78%. Recall scores show the model's ability to recollect relevant information: Descriptive questions 89%, Factual 87%, and Inferential 83%. With descriptive questions scoring 87%, factual questions 84%, and inferential questions 80%, the F1 Score shows good accuracy and memory. The boxplot and violin plot show the model's consistency across question categories, with descriptive questions having high median accuracy and precision.

Several ways to improve and explore this study arise as we envisage its future. First, increasing the dataset size might improve the model's knowledge and reaction. Transfer learning, pre-trained models, and innovative architectures may improve performance. The model's decision-making process's interpretability is fascinating for further study. Attention processes and visualization tools may reveal which picture areas and question components influence the model's replies. Furthermore, using the system for multimodal problems other than VQA is intriguing. Exploring real-world applications like picture captioning or visual dialogue may build model adaptability. This study provides a solid basis for VQA techniques in Multimodal Learning Frameworks, with future efforts to refine and expand the model's capabilities for multimodal applications.

A. Declaration Conflict of Interest

The authors declare that this manuscript has no conflict of interest with any other published source and has not been published previously (partly or in full). No data have been fabricated or manipulated to support our conclusions.

No funding is applicable and declaration for no financial Interest.

Acknowledgment

Acknowledgment the authors declare that they have no conflict of interest. The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript. The article has no research involving Human Participants and/or Animals. The author has no financial or proprietary interests in any material discussed in this article.

Compliance with Ethical Standards

Conflicts of Interest:

The authors declare that they have no conflict of interest. The manuscript was written through the contributions of all authors. All authors have approved the final version of the manuscript.

Availability of data and material:

Not data and materials are available for this paper. Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Ethical Approval:

The article has no research involving Human Participants and/or Animals

Competing Interest:

The author has no financial or proprietary interests in any material discussed in this article.

Declarations

Funding:

No Funding is applicable.

Code availability:

The data and code can be given based on the request

Consent to Participate:

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Consent to Publish:

All authors have given approval to the final version of the manuscript for publication.

References

[1] Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

[2] Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A. and Fidler, S., 2015. Skip-thought vectors. Advances in neural information processing systems, 28.

[3] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T., 2014, January. Decaf: A deep convolutional activation feature for generic visual recognition. In International conference on machine learning (pp. 647-655). PMLR.

[4] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D., 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

[6] Hodosh, M., Young, P. and Hockenmaier, J., 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47, pp.853-899.

[7] Li, S., Kulkarni, G., Berg, T., Berg, A. and Choi, Y., 2011, June. Composing simple image descriptions using web-scale n-grams. In

Proceedings of the fifteenth conference on computational natural language learning (pp. 220-228).

[8] Vedantam, R., Lawrence Zitnick, C. and Parikh, D., 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

[9] Winograd, T., 1972. Understanding natural language. Cognitive psychology, 3(1), pp.1-191.

[10] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

[11] Zhu, Y., Groth, O., Bernstein, M. and Fei-Fei, L., 2016. Visual7w: Grounded question answering in images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4995-5004).

[12] Yang, Z., He, X., Gao, J., Deng, L. and Smola, A., 2016. Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).

[13] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[14] Andreas, J., Rohrbach, M., Darrell, T. and Klein, D., 2016. Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705.

[15] Xiong, C., Merity, S. and Socher, R., 2016, June. Dynamic memory networks for visual and textual question answering. In International conference on machine learning (pp. 2397-2406). PMLR.

[16] Noh, H., Seo, P.H. and Han, B., 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 30-38).

[17] Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T. and Schiele, B., 2016. Grounding of textual phrases in images by reconstruction. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 817-834). Springer International Publishing.

[18] Hu, R., Rohrbach, M. and Darrell, T., 2016. Segmentation from natural language expressions. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 108-124). Springer International Publishing.

[19] Lin, T.Y., RoyChowdhury, A. and Maji, S., 2015. Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE international conference on computer vision (pp. 1449-1457).

[20] Gao, Y., Beijbom, O., Zhang, N. and Darrell, T., 2016. Compact bilinear pooling. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 317-326).

[21] Lu, J., Yang, J., Batra, D. and Parikh, D., 2016. Hierarchical co-attention for visual question answering. Advances in neural information processing systems (NIPS), 2.

[22] Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J., 2004. Canonical correlation analysis: An overview with application to learning methods. Neural computation, 16(12), pp.2639-2664.

[23] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J. and Lazebnik, S., 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13 (pp. 529-545). Springer International Publishing.

[24] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J. and Lazebnik, S., 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision (pp. 2641-2649).

[25] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A. and Mikolov, T., 2013. Devise: A deep visual-semantic embedding model. Advances in neural information processing systems, 26.

[26] Karpathy, A. and Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[27] Kiros, R., Salakhutdinov, R. and Zemel, R., 2014, June. Multimodal neural language models. In International conference on machine learning (pp. 595-603). PMLR.

[28] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuille, A., 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.

[29] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y., 2011. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 689-696).

[30] Malinowski, M. and Fritz, M., 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems, 27.

[31] Krishnamurthy, J. and Kollar, T., 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. Transactions of the Association for Computational Linguistics, 1, pp.193-206.

[32] Ren, M., Kiros, R. and Zemel, R., 2015. Exploring models and data for image question answering. Advances in neural information processing systems, 28.

[33] Ma, L., Lu, Z. and Li, H., 2016, March. Learning to answer questions from image using convolutional neural network. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

[34] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L. and Xu, W., 2015. Are you talking to a machine? dataset and methods for multilingual image question. Advances in neural information processing systems, 28.

[35] Karpathy, A., Joulin, A. and Fei-Fei, L.F., 2014. Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems, 27.

[36] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J. and Lazebnik, S., 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision (pp. 2641-2649).

[37] Karpathy, A. and Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[38] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[39] Andreas, J. and Klein, D., 2014, June. Grounding language with points and paths in continuous spaces. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning (pp. 58-67).

[40] Cadene, R., Ben-Younes, H., Cord, M. and Thome, N., 2019. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1989-1998).

[41] Li, L., Gan, Z., Cheng, Y. and Liu, J., 2019. Relation-aware graph attention network for visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10313-10322).

[42] Gao, L., Cao, L., Xu, X., Shao, J. and Song, J., 2020. Question-Led object attention for visual question answering. Neurocomputing, 391, pp.227-233.

[43] Sun, B., Yao, Z., Zhang, Y. and Yu, L., 2020. Local relation network with multilevel attention for visual question answering. Journal of Visual Communication and Image Representation, 73, p.102762.

[44] Zhang W, Jing Y, Hua H, Haiyang H, Qin Z (2020) Multimodal feature fusion by relational reasoning and attention for visual question answering. Information Fusion 55:116–126.

[45] Bai, Z., Li, Y., Woźniak, M., Zhou, M. and Li, D., 2021. DecomVQANet: Decomposing visual question answering deep network via tensor decomposition and regression. Pattern Recognition, 110, p.107538.

[46] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S. and Zhuang, Y., 2020. Counterfactual samples synthesizing for robust visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10800-10809).

[47] Sharma, H. and Jalal, A.S., 2021. Visual question answering model based on graph neural network and contextual attention. Image and Vision Computing, 110, p.104165.

[48] Lobry, S., Marcos, D., Murray, J. and Tuia, D., 2020. RSVQA: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 58(12), pp.8555-8566.

[49] Xi, Y., Zhang, Y., Ding, S. and Wan, S., 2020. Visual question answering model based on visual relationship detection. Signal Processing: Image Communication, 80, p.115648.

[50] Basu, K., Shakerin, F. and Gupta, G., 2020, January. Aqua: Asp-based visual question answering. In International Symposium on Practical Aspects of Declarative Languages (pp. 57-72). Cham: Springer International Publishing.

[51] Sharma, H. and Jalal, A.S., 2020. Incorporating external knowledge for image captioning using CNN and LSTM. Modern Physics Letters B, 34(28), p.2050315.

[52] Geman, D., Geman, S., Hallonquist, N. and Younes, L., 2015. Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences, 112(12), pp.3618-3623.

[53] Das, A., Kottur, S., Moura, J.M., Lee, S. and Batra, D., 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In Proceedings of the IEEE international conference on computer vision (pp. 2951-2960).

[54] De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H. and Courville, A., 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5503-5512).

[55] Seo, P.H., Lehrmann, A., Han, B. and Sigal, L., 2017. Visual reference resolution using attention memory for visual dialog. Advances in neural information processing systems, 30.

[56] Massiceti, D., Siddharth, N., Dokania, P.K. and Torr, P.H., 2018. Flipdial: A generative model for two-way visual dialogue. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6097-6105).

[57] Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K. and Darrell, T., 2016. Natural language object retrieval. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4555-4564).

[58] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L. and Murphy, K., 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 11-20).

[59] Yu, L., Poirson, P., Yang, S., Berg, A.C. and Berg, T.L., 2016. Modeling context in referring expressions. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14 (pp. 69-85). Springer International Publishing.