IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141219

A Review on Visual Question Answering By Image Captioning

Sarah Jose[1], Goutham Krishna L U[2]

Student, MSc Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India[1]

Assistant Professor, PG Department of Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India[2]

Abstract: Visual Question Answering (VQA) is a complex multimodal task that requires instant understanding of visual content and natural language queries, yet traditional models often struggle to construct a complete semantic representation of the same. Although conventional VQA systems rely on deep visual feature extraction and linguistic encoders for the question, they commonly fail to capture global context, exact object interactions, and long-range dependencies. A major limitation across early VQA models is the presence of strong language bias, where the system predicts answers based on frequently occurring question-answer patterns rather than genuine visual grounding. To address these issues, recent research has introduced image captioning as a complementary semantic modality capable of enriching scene understanding. Captions provide descriptive information about object attributes, relationships, and contextual cues that may be missing or underrepresented in raw visual features, and integrating them through attention mechanisms such as Attention Aware modules or Question-Guided Parallel Attention allows models to filter irrelevant tokens and retain meaningful semantics. This fused representation creates a more robust and contextually aligned multimodal embedding that strengthens reasoning across diverse question types. Experimental results on benchmark datasets show that caption-enhanced approaches offer consistent improvements in accuracy and interpretability, although they remain dependent on caption quality and introduce additional computational complexity. Nonetheless, the integration of caption-generated semantics represents a promising direction toward developing more context-aware and visually grounded VQA systems capable of more reliable and human-like reasoning.

Keywords: Visual Question Answering, Attention Aware, Question-Guided Parallel Attention, Image Captioning, Deep Learning, VQA v1,VQA v2.

I. INTRODUCTION

Visual Question Answering is a rapidly evolving area of artificial intelligence that has gained attention due to its ability to mimic aspects of human multimodal reasoning. In a typical VQA task, the system must understand an image, interpret a natural language question about that image, and produce a relevant answer. However, this seemingly straightforward process involves numerous layers of complexity, including object recognition, spatial understanding, linguistic interpretation, and contextual reasoning. Traditional VQA models attempt to solve these challenges by combining convolutional image encoders with recurrent or transformer-based question encoders, but these methods are often limited in their ability to capture the full meaning of the visual scene.

One of the primary limitations of standard VQA models is their reliance on object-level features extracted from region proposals or bounding box detectors. While such features are useful for identifying individual objects, they often fail to capture interactions or relationships between objects, such as actions, spatial arrangements, and context. For example, identifying a person and a bicycle is not enough to answer the question “What is the person doing?” unless the model understands that the person is riding the bicycle. This lack of semantic completeness frequently leads to incorrect or incomplete reasoning, especially for questions requiring inferential thinking.

In addition to incomplete visual understanding, many VQA models suffer from strong language bias. Large-scale VQA datasets often contain repetitive question-answer patterns, which the model can exploit to guess answers without adequately analyzing the image. For instance, yes/no questions are often skewed toward "yes" and common color questions may default to "blue" This bias allows models to achieve artificially inflated accuracy scores without genuinely understanding the image. As a result, the predictions become unreliable in real-world scenarios where visual content may differ from the patterns seen in the training dataset.

DOI: 10.17148/IJARCCE.2025.141219

To address these challenges, researchers have explored integrating image captioning into the VQA pipeline as a means of providing enhanced semantic context. Captions present a rich, descriptive summary of the scene, offering information that visual features alone may fail to capture. When used effectively, captions can help the model better understand actions, relationships, and global scene structure, enabling deeper reasoning and reducing reliance on dataset biases. This shift toward using captions marks an important evolution in VQA research, bringing models closer to human-like visual understanding.

II. BACKGROUND AND CONTEXT

The foundations of VQA research lie in the broader fields of computer vision and natural language processing. Early models treated these tasks separately, with vision networks extracting spatial or object-level features and language networks transforming the question into a vector representation. The fusion of these modalities was often performed through simple concatenation or element-wise operations, which limited the model’s ability to perform complex interactions between the two feature types. These simplistic fusion approaches constrained the model’s reasoning capacity and often missed important relationships required to answer more complex questions.

As research progressed, attention mechanisms became a transformative technique in VQA. Attention allowed models to focus selectively on the parts of the image most relevant to the question, introducing interpretability and improving accuracy. Models implementing co-attention mechanisms further advanced this capability by enabling simultaneous alignment between question words and image regions. However, even attention-based models struggled to capture global scene context, as they were still fundamentally limited by the visual features fed into them. Attention could highlight important regions, but it could not compensate for missing semantic information.

The parallel evolution of image captioning models offered a new source of semantic information that could benefit VQA. Captioning models generate coherent sentences describing the image, effectively translating visual content into a structured linguistic form. These descriptions often include high-level semantic relationships, such as the action being performed, the spatial arrangement of objects, and contextual background details. These elements are particularly beneficial for VQA tasks involving reasoning, inference, or understanding of object interactions. Captioning thus emerged as a valuable supplementary modality that could enrich VQA systems with deeper semantic context.

However, incorporating captions into VQA presents specific challenges. Not all caption words are useful for answering a particular question, and unfiltered captions may introduce noise or irrelevant details. This makes selective filtering essential for effective caption integration. Furthermore, merging three modalities image, question, and caption requires advanced fusion techniques capable of balancing and aligning semantic information across each modality. These challenges have motivated the development of specialized attention modules and multimodal fusion strategies designed to harness captions effectively while avoiding common pitfalls.

III. RELATED WORKS

Recent developments in Visual Question Answering (VQA) have increasingly explored the use of image captioning to strengthen semantic understanding and reduce language bias. Since traditional models often struggle with complex scene interpretation, several studies have proposed caption-guided attention and fusion techniques to improve reasoning accuracy. The following works summarize key contributions in this area and highlight how caption-enhanced approaches advance VQA performance.

Improving Visual Question Answering by Image Captioning[1] proposes a caption-guided VQA framework that integrates captions with visual and question features using attention modules for cleaner and more effective fusion.

VQA: Visual Question Answering[2] introduced the first large-scale VQA dataset and baseline models, enabling standardized evaluation but suffering from language priors and limited reasoning.

Stacked Attention Networks[3] uses multiple stacked attention layers to iteratively focus on image regions based on the question, improving accuracy but struggling with multi-step reasoning.

A Strong Baseline for VQA[4] provides a simple but competitive ResNet + GRU/LSTM attention model that is fast and reproducible but lacks deeper reasoning ability.

DOI: 10.17148/IJARCCE.2025.141219

Making the V in VQA Matter (VQA v2.0)[5] reduces language bias by pairing balanced question-image examples, improving vision-grounded reasoning while still leaving many challenges unresolved.

Multimodal Compact Bilinear Pooling[6] efficiently approximates bilinear interactions between vision and language features to boost VQA accuracy, though at higher computational cost.

Bilinear Attention Networks (BAN)[7] models fine-grained bilinear interactions between image regions and question words for strong accuracy but with heavy memory and compute demands.

MCAN: Deep Modular Co-Attention Networks[8] uses stacked co-attention layers to tightly align image and question representations, achieving high performance but with increased complexity.

Bottom-Up and Top-Down Attention[9] combines object-level region features with top-down attention to improve grounding and accuracy, relying heavily on pretrained object detectors.

RAMEN[10] aggregates multimodal embeddings recurrently to generalize well across datasets but lacks the relational reasoning capability of transformer models.

Hierarchical Question-Image Co-Attention[11] attends jointly to multiple levels of question structure and image regions, improving interpretability but limited by older CNN/LSTM components.

MUREL[12] uses relational reasoning cells to model rich region-to-region interactions for deeper reasoning, though at significant computational cost.

ViLBERT[13] employs dual-stream transformers with co-attention for strong multimodal representations but is computationally intensive.

LXMERT[14] uses separate vision, language and cross-modality transformers pretrained on multiple tasks to enable complex reasoning but requires large-scale training.

BUTD + GloVe Baseline[15] combines object-region features with simple GloVe text embeddings to form an easy baseline that performs reasonably but lacks fine-grained multimodal alignment.

IV. SYSTEMATIC ANALYSIS

The systematic analysis provides a clear and organized examination of the model by breaking down its key components, performance measures, and functional behavior. Since the approach combines image features, question information, and caption-based semantics, it is important to present these elements in a structured manner to understand how each contributes to the overall performance. The following table summarizes the model's accuracy, methodology, advantages, disadvantages, and datasets used, offering a straightforward overview of the system's strengths and limitations.

Ref no.	Accuracy	Advantages	Disadvantages	Methodology	Dataset
[1] Shao et al.	72.55%	Combines image caption semantics with visual features, improves reasoning and answer accuracy, interpretable.	Weaker on number questions, depends on caption quality, needs more computation.	Uses Attention Aware (AA) module to filter caption noise and Question Guided Parallel Attention (QGPA) to fuse image, caption, and question features.	VQA v1.0, VQA v2.0, COCO captions
[2] Antol et al.	54.1%	First large scale VQA dataset, strong baseline model combining vision and language.	Strong language bias, limited reasoning ability.	CNN extracts image features, LSTM encodes question, then both fused to predict answers.	VQA v1.0

DOI: 10.17148/IJARCCE.2025.141219

[3] Yang et al.	57.6%	Multiple attention layers refine focus, interpretable attention visualization.	Struggles with multi step logical reasoning, limited to attention over image only.	Stacked attention layers applied to visual features guided by the question embedding.	VQA v1.0
[4] Kazemi & Elqursh	64.6%	Simple, fast to train, and competitive baseline, reproducible model.	Can't handle reasoning or compositional questions, limited generalization.	Uses ResNet features for images and GRU for question encoding with soft attention.	VQA v1.0
[5] Agrawal et al.	63%	Reduces dataset bias by using balanced image question pairs.	Still limited for reasoning and relational understanding.	Creates paired datasets and trains models to look at visual evidence instead of priors.	VQA v2.0
[6] Fukui et al.	59.8%	Powerful multimodal fusion method, better representation of image and text.	Requires more computation and memory, sensitive to Hyperparameters.	Combines CNN image and LSTM question features via compact bilinear pooling and attention.	VQA v1.0
[7] Kim et al.	70%	Models fine grained interactions between words and image regions, improves accuracy.	High Computational cost, needs large GPU memory.	Bilinear attention maps between image regions and question words, fused features classify answer.	VQA v2.0
[8] Yu et al.	70.63%	Strong state of the art model, deep co attention captures question to image and image to question relations.	Slower training due to deep layers, more parameters.	Stacked co attention modules with residual connections between question and image features.	VQA v2.0
[9] Anderson et al.	70%	Introduced region based (object level) attention, improved visual grounding.	Needs pre trained object detectors, more preprocessing steps.	Bottom up: Faster R CNN proposes regions, Top down: task driven attention selects relevant regions.	COCO (for detector), VQA v2.0
[10] Shrestha et al.	62%	Strong performance for simple architecture, robust cross-dataset generalization.	Lower accuracy than transformer models, limited relational reasoning.	Recurrent Aggregation of Multimodal Embeddings combining simple feature fusion and aggregation.	VQA v2.0, CLEVR, TDIUC
[11] Lu et al.	57%	Multi-level question/image co-attention, interpretable	Weak on multi-step reasoning, older CNN/RNN	Hierarchical co-attention over words, phrases, and full question	VQA v1/v2

[12] Cadene et al.	65%	Strong relational reasoning, handles object interactions	More parameters, slower training	Relation based fusion between detected objects	VQA v2, Visual Genome
[13] Lu et al.	71%	Two-stream Transformer, powerful cross modal alignment	Heavy pretraining, large compute	Separate visual & text streams with co-attention	VQA v2, COCO, VG
[14] Tan & Bansal	72%	Strong cross-modal encoder, robust across tasks	High computation, large model	Language encoder + object encoder + cross-modal Transformer	VQA v2, COCO, VG
[15] Anderson et al.	59%	Simple, fast baseline, easy to train, uses strong region features.	Low accuracy, lacks fine-grained alignment, no advanced attention.	Bottom-Up Faster R-CNN features + GloVe embeddings + simple classifier.	VQA v2

V. CONCLUSION AND FUTURE WORK

Caption-assisted VQA represents an important advancement in the field of multimodal reasoning, offering a richer and more context-aware approach to understanding images and answering questions. By incorporating captions into the reasoning pipeline, models benefit from an additional layer of semantic interpretation that complements object-level visual features. This added context helps reduce common issues such as language bias and incomplete scene understanding, enabling more accurate and robust answer generation. Through careful filtering of caption tokens and sophisticated multimodal fusion techniques, these systems achieve deeper alignment between image, caption, and question modalities.

The gains provided by caption integration come with certain challenges, including increased computational demands and reliance on caption quality. Future work may explore more efficient captioning models, perhaps leveraging transformers or large-scale vision-language architectures that generate richer and more accurate descriptions. Further research could also investigate methods for generating multiple captions per image to provide diverse semantic perspectives, improving the overall robustness of the system. Advances in noise-filtering mechanisms and cross-modal alignment strategies may also help reduce the impact of irrelevant or misleading caption components.

Another promising direction is the incorporation of real-world constraints, such as real-time processing and memory efficiency. Developing lightweight captioning and VQA models suitable for deployment on mobile or embedded devices could significantly broaden the applicability of this technology. Beyond technical improvements, integrating external knowledge sources such as knowledge graphs or commonsense reasoning models may allow VQA systems to answer more complex or inferential questions that require understanding beyond the visible scene.

REFERENCES

[1] X. Shao et al., "Improving Visual Question Answering by Image Captioning," Conference on Visual Question Answering, 2025.
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "VQA: Visual Question Answering," IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2425-2433.
[3] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked Attention Networks for Image Question Answering," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29.
[4] V. Kazemi and A. Elqursh, "A Strong Baseline for Visual Question Answering," European Conference on Computer Vision Workshops (ECCVW), 2017, pp. 1-6.
[5] A. Agrawal, D. Batra, D. Parikh and C. L. Zitnick, "Making the V in VQA Matter: VQA 2.0," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6904-6913.
[6] A. Fukui et al., "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 457-468.

DOI: 10.17148/IJARCCE.2025.141219

[7] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha and B.-T. Zhang, "Bilinear Attention Networks," Neural Information Processing Systems (NeurIPS), 2018, pp. 1571-1581.
[8] T. Yu, J. Yu, Y. Cui, D. Tao and Q. Tian, "Deep Modular Co-Attention Networks for Visual Question Answering," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6281-6290.
[9] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and VQA," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6077-6086.
[10] R. Shrestha, K. Kafle and C. Kanan, "Recurrent Aggregation of Multimodal Embeddings (RAMEN)," IEEE Winter Conference on Applications of Computer Vision (WACV), 2019, pp. 1-10.
[11] J. Lu, J. Yang, D. Batra and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," Neural Information Processing Systems (NeurIPS), 2016, pp. 289-297.
[12] R. Cadene, H. Ben-Younes, M. Cord and N. Thome, "MUREL: Multimodal Relational Reasoning for Visual Question Answering," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1989-1998.
[13] J. Lu, D. Batra, D. Parikh and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," Neural Information Processing Systems (NeurIPS), 2019, pp. 13-23.
[14] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations for Vision-and-Language Tasks," Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 5099-5111.
[15] P. Anderson et al., "BUTD + GloVe Baseline for Visual Question Answering," VQA Benchmark Reports, 2018.