Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34229

Progress and Analysis of Visual Question Answering from Traditional Methods to Multimodal Large Models

Xiaohan Liu

School of Information and Communication Engineering, Hainan University, Haikou, China
nancyxiaohan@outlook.com

Abstract. Visual Question Answering (VQA) is a multimodal task in field of artificial intelligence. It combines computer vision and natural language processing. System can understand image content and combine it with question semantics to generate correct answers. This is important benchmark for testing machine cross-modal understanding ability. This paper give a full review of VQA, focusing on its definition, research methods and future trends. First, this paper introduce definition of VQA tasks and the question-answer format. Then, divide VQA into three stages according to technology evolution: traditional deep learning, Transformer-based vision-language pretraining, and the stage using knowledge-enhanced generative reasoning with multimodal large models. Based on this, this paper summarize common datasets and evaluation metrics. It is worth noting that also analyze existing problems and discuss future development directions. This work can provide technical roadmap for future VQA research and can give theoretical guidance for building robust VQA models with good generalization.

Keywords: Visual Question Answering, Deep Learning Methods, Transformer Methods, Multimodal Large Model Methods

1. Introduction

Enabling machines to 'see' images and 'understand' language is a long-term goal of artificial intelligence. VQA is a key task for achieving this goal: it requires systems to answer natural language questions based on understanding visual content, lying at the intersection of computer vision and natural language processing. It is not only an important touchstone for testing the comprehensive intelligence of machines but also has broad application prospects in several areas, including assistance for the visually impaired, medical diagnosis, intelligent education, and human-computer interaction [1].

The core challenge of VQA lies in its cross-modal nature: systems must understand visual semantics and linguistic intent, integrating them to generate accurate answers [2]. In recent years, rapid development in deep learning has driven VQA through three major technological leaps: traditional deep learning with joint embedding and attention mechanisms; Transformer-based vision-language pretraining that learns universal cross-modal representations from large-scale image-text pairs; and multimodal large model-based knowledge-enhanced reasoning, using large language

DOI: 10.54254/2755-2721/2026.GU34229

models as the core engine to achieve deep vision-language fusion with strong zero-shot and reasoning capabilities.

This paper provides a systematic review of the development trajectory of VQA technology, aiming to clarify the technical paradigms, core breakthroughs, and evolution logic of each stage, summarize commonly used datasets and evaluation metrics, analyze the core challenges currently faced such as multimodal hallucinations and limitations in compositional reasoning, and prospect future development directions. It seeks to offer a clear technical roadmap for subsequent research in the VQA field and to provide theoretical support and practical guidance for advancing cross-modal understanding technologies towards general artificial intelligence.

2. Definition and types of VQA

The VQA task can be formally defined as: given visual input V and a natural language question Q, the system generates a natural language answer A. That is, $A=f(V,Q;\theta)$, where f represents a model with parameters θ [3]. The visual input V can take various forms, such as a single image, video, chart, document, etc. The question Q is a free-form natural language question, and the answer A can be a predefined category or free text [4].

According to the way answers are generated, VQA can be divided into open-ended questions and multiple-choice questions [3]. According to knowledge requirements, it can be divided into purely visual VQA and knowledge-based VQA [5]. The formal representation of knowledge-based VQA is: given an image V , a question Q, and an external knowledge base K, the system needs to generate an answer $A=f(V, Q, K;\theta)$ [5].

3. Evolution of VQA methods

3.1. Visual question answering based on traditional deep learning methods

Traditional deep learning methods represent the early exploration of VQA, characterized by the use of a "encode-fuse-classify" joint embedding paradigm. The advantage of these methods lies in their simple structure and ease of training, but the drawback is that they are limited by global image features, making it difficult to precisely locate local areas relevant to the question [6].

3.1.1. Methods

Joint Embedding Framework. The VIS LSTM method proposed by Ren et al. pioneered modeling VQA as a classification problem. This method first uses a pre-trained CNN to extract a global feature vector of the entire image, while encoding the variable-length question into a fixed-dimensional semantic vector using an LSTM. Then, the image and question feature vectors are simply concatenated or element-wise multiplied, and finally the result is input into a Softmax classifier to select the most probable answer from a predefined answer set [4].

Attention Mechanism. Shih et al. proposed the Region Selection Network to address global feature limitations: the Edge Boxes algorithm extracts about 100 candidate image regions, a CNN extracts their visual features, the model calculates relevance scores via the inner product of region and question features, converts these to attention weights with Softmax, and aggregates weighted regional features to get a question-relevant feature vector for answer prediction [7].

DOI: 10.54254/2755-2721/2026.GU34229

3.1.2. Applications

Assistive Technology to Help the Visually Impaired. In fact, traditional deep learning methods are usually lightweight and efficient. So they can be used in mobile deployment for real-time assistive applications to help the visually impaired people, which means can run the algorithm on phones and other devices with limited computing power and this is important for practical use. Early attention-based assistive systems can address problems such as blurred images and random angles by using CNNs for feature extraction and LSTMs for encoding voice questions, while the region-selection attention mechanism is expected to focus on key areas in the image so that the system can work better [7]. Then use a classifier to generate answers. This can enable real-time voice interaction after image capture is finished, so that users can get voice feedback immediately [8].

3.2. Visual question answering based on transformer and pre-training

Transformers and pretraining bring a new way to VQA. They use self-attention to learn general cross-modal features from a large number of image-text pairs, which are widely used in current research. This method can capture long-range connections and has good reasoning ability. However, it needs a lot of computing resources and data to get good results. This is because the model is usually large, and its training takes a long time to finish [8, 9]. Salemi et al. further enhanced this paradigm by pretraining dense retrieval models via contrastive learning to provide factual support for VQA [10].

3.2.1. Methods

ViLBERT (dual-stream architecture) Lu et al. proposed ViLBERT [11]. In fact, it is an early work in visual-linguistic pretraining and uses dual-stream architecture. The model has two Transformer streams. Visual stream and language stream are separate. Visual stream can use Faster R-CNN to get region features from images and language stream uses BERT to encode the text and then two streams interact through co-attention layers so that cross-modal features can be improved with bidirectional guidance. In pretraining stage, ViLBERT is trained with masked multimodal modeling and multimodal alignment prediction on Conceptual Captions dataset. After it learns general cross-modal representations, and fine-tune it on downstream tasks such as VQA v2 and need to add a classification layer because this can help the model adapt to different tasks without changing the main structure.

UNITER (Single-Stream Architecture) Chen et al. proposed UNITER. It is one representative of single-stream architecture. Different from ViLBERT, UNITER puts image regional features and text tokens into one sequence. Then input them to a unified Transformer to encode together. In pre-training stage, use several tasks. In fact, the model handles mask language modeling and mask area modeling. Image-text matching is also used. These can help it learn better cross-modal alignment [8].

3.2.2. Applications

Remote sensing image analysis. Transformers and pre-training methods have cross-modal alignment capabilities, so in principle they should be suitable for remote sensing image analysis, which is a scenario that requires handling large amounts of data and complex relationships between images and text at the same time. They use VQA-AID model from Sarkar et al. It applies this approach to post-

DOI: 10.54254/2755-2721/2026.GU34229

disaster damage assessment and performs question-answer analysis on remote sensing images by asking questions such as "Are the buildings damaged?" and this process can help us to identify the affected areas that should be prioritized. The process includes inputting post-disaster remote sensing images. And extract visual features through pre-trained model and generate answers and assess the severity of the disaster based on these results so that the final evaluation can be obtained. In fact, this method can be used in post-disaster emergency response and it helps improve assessment efficiency [8].

3.3. Visual question answering based on multimodal large model methods

Multimodal large models can be seen as main trend in VQA now. In fact, they use LLMs' reasoning ability and world knowledge, and combine these with visual models through freezing or fine-tuning to build the full system. The advantage is that it has good zero-shot generalization and can handle complex reasoning. But there are also problems. The deployment cost is high. Also, inference latency is large and hallucinations can happen.

3.3.1. Methods

Freeze the LLM paradigm . Prophet method from Shao et al. is one freeze LLM paradigm. In fact, first train a small VQA model. It can generate answer candidates and also answer-aware examples as heuristics. Then this heuristic information, together with image description and question, is used to build complex text prompts and feed into GPT-3 model with frozen parameters. GPT-3 needs no parameter update. It can use contextual learning ability to infer the final answer from prompts [12].

LLaVA (Instruction Fine-Tuning Paradigm). Liu et al. proposed LLaVA as method of instruction fine-tuning. And use pre-trained CLIP vision encoder to extract image features, and then a simple learnable projection layer is used to map these visual features to embedding space of large language models so that the alignment between vision and language can be achieved and the model can understand visual content. Then the projection layer and LLM are fine-tuned end-to-end on large-scale generated vision instruction data. It is worth noting that LLaVA can show strong multimodal dialogue and reasoning capabilities and it marks the evolution of VQA from simple question answering to general visual understanding and dialogue [13].

3.3.2. Applications

Medical diagnostic aid. In fact, medical diagnosis requires high accuracy and good interpretability, and should also integrate professional knowledge in system because the clinical environment can be complex and can have many uncertainties. Multimodal large models can use knowledge reasoning and generate detailed explanations, so they can be an ideal solution in this field. Medical VQA systems are one typical application can look at. The VQA-RAD dataset proposed by Lau et al. includes 315 radiology images and 3,515 question-answer pairs generated by clinicians, and this dataset is used to test whether the model can understand medical content correctly under different conditions. Gong et al. use multitask pretraining and cross-modal self-attention. Their method achieved 73.2% accuracy on this dataset which shows it can work in principle. It can answer questions such as 'Is there an abnormality in the lungs?' [14]. Recently, medical multimodal large models such as LLaVA-Med proposed by Li et al. can answer clinical questions based on medical images, and it can help reduce the burden on doctors and provide decision support so that the diagnosis process can be improved [15].

DOI: 10.54254/2755-2721/2026.GU34229

4. Common VQA datasets and evaluation metrics

4.1. Major datasets

4.1.1. General datasets

VQA v1 from Antol et al. [3] contains 204,721 images and 614,163 question-answer pairs and this is the first large-scale VQA dataset which can be used for various tasks in computer vision and natural language processing so it is very important for the community. Then, VQA v2 from Ishmam et al. [8] introduces complementary image pairs which can reduce language bias and it expands scale to 1,105,904 question-answer pairs. GQA [8] is proposed by Hudson and Manning. It has 113,000 images and 22 million question-answer pairs. In fact, people can use this dataset to evaluate compositional reasoning abilities and the answer distribution is balanced.

4.1.2. Diagnostic datasets

CLEVR [4] is from Johnson et al. It contains 70,000 synthetic 3D images and about 700,000 questions. People can use this dataset to evaluate compositional reasoning and multi-step reasoning abilities.In fact, VQA-CP [8] is proposed by Agrawal et al. They re-split the VQA v2 training and testing sets, and the answer distribution changes. It can be used to evaluate whether a model can overcome language priors.WHOOPIS [8] is proposed by Bitton-Guetta et al. It uses the AI-generated images with unreasonable or counter-intuitive content. This dataset can test a model's reasoning ability and commonsense understanding.

4.1.3. Knowledge-based datasets

OK-VQA [12] is from Marino et al. and it contains 14,031 images. Also there are 14,055 questions which can require external knowledge to answer. It covers 10 knowledge categories. A-OKVQA [12] is from Schwenk et al. and it contains about 25,000 questions. Each is accompanied by reasoning evidence. So it can be more difficult than OK-VQA.FVQA [5] is from Wang et al. It contains about 4,600 questions, each providing supporting facts and involving 580 visual concepts.

4.2. Evaluation metrics

VQA accuracy is defined as receiving full score if the predicted answer matches the answers of at least 3 annotators [3]:

$$Accuracy = min\left(\frac{N_{match}}{3}, 1\right) \tag{1}$$

where Nmatch represents number of annotators giving this answer (that is, number of human annotators whose answers match the model's predicted answer) and min(_, 1) takes smaller value between calculation in parentheses and 1, so accuracy can avoid exceeding 1.

MPT refers to the mean accuracy across categories, compensating for imbalanced class distribution [16]:

$$\mathrm{MPT} = \frac{1}{\mathrm{C}} \sum_{\mathrm{i}=1}^{\mathrm{C}} \mathrm{Acc_i} \tag{2}$$

DOI: 10.54254/2755-2721/2026.GU34229

where C is the total number of question categories, and Acci is the accuracy for the i-th question category.

MM-Vet is a multidimensional comprehensive evaluation benchmark covering six core abilities including visual perception, knowledge integration, and logical reasoning [13].

Hallucination rate is used to evaluate the proportion of model-generated answers that do not match the visual content, reaching as high as 15-30% in open-ended generation tasks [13].

5. Challenges and future prospects

5.1. Current major challenges

Current VQA and multimodal models have several problems. They can suffer from multimodal hallucination and compositional reasoning limitations, and there are also efficiency-scalability trade-offs, reliance on statistical shortcuts, plus multilingual imbalance. Models can generate answers that do not match visual content, and it is worth noting that hallucination rates can reach 15 - 30% in open-ended generation tasks [13], while they can struggle to generalize to unseen concepts and reasoning combinations when facing new visual patterns in the test data. Performance can drop when reasoning depth increases [13]. Large models have high inference latency. Also, the memory consumption is large, which can restrict deployment of real-time applications [6]. Models can rely on statistical shortcuts rather than true visual understanding, leading to a decline in performance in out-of-distribution scenarios where the training distribution and test distribution are different [2]. In fact, existing research is centered on English, and there are shortcomings in cross-lingual and cross-cultural visual understanding abilities [13].

5.2. Future development directions

Future work should focus on several directions. Hallucination can be controlled through self-reflective architecture. People can also use visual grounding for uncertainty quantification. Neural-symbolic reasoning transforms visual reasoning into executable symbolic programs. This can enhance compositional generalization. People should build unified efficiency architectures that integrate parameter-efficient fine-tuning, model compression, and the inference optimization. It is worth noting that causal evaluation systems using counterfactual and adversarial testing are expected to help distinguish genuine understanding from statistical shortcuts. Finally, cross-cultural multilingual VQA datasets with culturally aware vision-language alignment methods should be developed.

6. Conclusion

VQA technology has three main stages in its development, and if look at the whole history, it can move from shallow feature extraction to deep neural architectures, and from specialized single tasks to generalized multimodal applications that can handle diverse scenarios. The stages include traditional deep learning, Transformer pre-training, and multimodal large models using LLMs as the core. In traditional deep learning stage, use joint embedding framework and CNNs and LSTMs encode images and questions with separate encoders. This establishes the basic paradigm of encode-fuse-classify. Then the Transformer and pre-training stage comes. It introduces self-attention mechanisms, and people can learn cross-modal representations from large-scale image-text pairs through pre-training, which is expected to enhance the model's generalization capability. In the third

DOI: 10.54254/2755-2721/2026.GU34229

stage, large language models become the core engine. It is worth noting that VQA can shift from simple perception tasks to complex cognitive reasoning through knowledge-enhanced methods or instruction fine-tuning, and this change can support more advanced applications in real-world scenarios.

Because of these technologies, VQA now can work in different field. It can help blind people in the daily life and it is used in medical diagnosis. Also, smart education is another important field. But in fact, current systems still have some problems people should note. They can produce multimodal hallucinations when processing complex inputs. Also, the compositional reasoning ability is limited because the model cannot understand complex relationships between objects. The system may depend on statistical shortcuts when making predictions based on training data biases instead of real understanding, and it cannot generalize to different languages which is a big barrier for practical use in non-English speaking countries and limits its global application. In the future, people should solve these problems. People can work on hallucination control and neuro-symbolic reasoning to improve the reasoning ability so that the model can think step by step. Also, unified architectures and cross-cultural alignment are important. This can help VQA move from just answering questions to become a trustworthy system with universal visual intelligence that can work in different cultures and languages without producing wrong information.

References

[1] Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 163, 3–20. https: //doi.org/10.1016/j.cviu.2017.06.005
[2] Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models.
[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2425–2433).
[4] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 163, 21–40.
[5] Ding, Y., Yu, J., Liu, B., Hu, Y., Cui, M., & Wu, Q. (2022). MuKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5089–5098).
[6] Teney, D., Anderson, P., He, X., & van den Hengel, A. (2018). Tips and tricks for visual question answering: Learnings from the 2017 challenge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4223–4232).
[7] Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4613–4621).
[8] Ishmam, M. F., Shovon, M. S. H., Mridha, M. F., & Dey, N. (2024). From image to language: A critical analysis of visual question answering (VQA) approaches, challenges, and opportunities. Information Fusion, 106, 102270. https: //doi.org/10.1016/j.inffus.2024.102270
[9] Zhou, Y., Ren, T., Zhu, C., Sun, X., Liu, J., Ding, X., Xu, M., & Ji, R. (2021). TRAR: Routing the attention spans in transformer for visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 2074–2084).
[10] Salemi, A., Rafiee, M., & Zamani, H. (2023). Pre-training multi-modal dense retrievers for outside-knowledge visual question answering. In Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR) (pp. 169–176). https: //doi.org/10.1145/3578337.3605137
[11] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic vision-language representations for vision-and-language tasks. In Advances in Neural Information Processing Systems (NeurIPS), 32.
[12] Shao, Z., Yu, Z., Wang, M., & Yu, J. (2023). Prompting large language models with answer heuristics for knowledge-based visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14974–14983).
[13] Riskin, E. (2026). Multimodal large language models for visual question answering: A comprehensive survey. Preprint. https: //doi.org/10.13140/RG.2.2.13150.50245

DOI: 10.54254/2755-2721/2026.GU34229

[14] Gong, H., Chen, G., Liu, S., Yu, Y., & Li, G. (2021). Cross-modal self-attention with multi-task pre-training for medical visual question answering. In Proceedings of the International Conference on Multimedia Retrieval (ICMR) (pp. 456–460). https: //doi.org/10.1145/3460426.3463584
[15] Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., & Ge, Z. (2023). Medical visual question answering: A survey. Artificial Intelligence in Medicine, 143, 102611. https: //doi.org/10.1016/j.artmed.2023.102611
[16] Kafle, K., & Kanan, C. (2017). An analysis of visual question answering algorithms. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 1965–1973)