Multimodal Question Answering: Method Evolution, Challenges and Prospects

Haopeng Li

Information Science and Technology College, Dalian Maritime University, Dalian, Liaoning, China
gbp5003@gmail.com

Abstract. With the breakthroughs in cross-modal technology of artificial intelligence, multi-modal question answering (MMQA), as a key research direction connecting image, text and voice information, has increasingly significant application value in fields such as barrier-free services and education. This paper systematically re-views from three dimensions: task classification, core model methods and experimental performance, and focuses on analyzing the technical paths of typical tasks such as visual question answering, voice VQA and image-to-voice QA. It also summarizes the innovative mechanisms and practical effects of pre-trained models like BLIP-2 and GIT in cross-modal representation and semantic under-standing. By comparing the dataset adaptability and evaluation results of different tasks, the study reveals the bottlenecks in the current technology in terms of se-mantic alignment quality and output naturalness. This analysis provides theoretical and practical references for the technical optimization and scenario expansion in the MMQA field. Future research can focus on improving the deep alignment of cross-modal semantics and the naturalness of generation, while exploring more general and adaptable frameworks to promote the in-depth application and innovative development of MMQA technology in complex real-world scenarios.

Keywords: Multimodal Question Answering, Visual Question Answering, Cross-modal Fusion, Pre-trained Model.

1 Introduction

The iterative development of artificial intelligence technology has driven the demand for machines to process heterogeneous information. The single-modal information interaction has been unable to meet the application expectations in complex scenarios. Multi-modal question answering (MMQA) integrates the visual features of images, the semantic information of text, and the auditory signals of speech to achieve precise responses to cross-modal queries. This capability provides core technical support for scenarios such as visual impairment user assistance systems and intelligent teaching tools [1].

The current development in the MMQA field relies on the breakthroughs in visual-language pre-training frameworks. Models such as BLIP-2 and Flamingo have

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_38

significantly improved task processing accuracy through innovative cross-modal alignment mechanisms [2]. However, existing research mostly focuses on optimizing a single task and lacks a systematic review of tasks involving different modalities. Moreover, the summaries of technical limitations have not yet formed a coherent framework.

This research takes the task types of MMQA as the core framework, and successively analyzes the technical logic of visual question answering, voice VQA, and image-to-speech QA. It combines literature cases to summarize the method characteristics and performance. Through comparative experimental data, it clarifies the differences in model adaptability and analyzes the current technical bottlenecks and future directions. This research aims to integrate scattered technical achievements and provide a structured reference for subsequent innovations in the field.

2 Visual Question Answering

Visual Question Answering (VQA), as a fundamental task type of Multi-modal Question Answering (MMQA), takes images and text questions as inputs and outputs natural language answers. The core challenge of this task lies in achieving precise alignment between visual features and language semantics [3]. This task is widely applied in scenarios such as image content interpretation and intelligent security, and is a key benchmark for evaluating cross-modal understanding capabilities.

The BLIP-2 model proposed by Li et al. adopts a two-layer architecture of "frozen image encoder and large language model (LLM)", and achieves cross-modal information transmission through a lightweight bridge module [4]. This model does not conduct secondary training on the image encoder in the VQA task, but only learns the mapping relationship between visual features and the language space through the bridge module, significantly reducing the training cost. In the zero-shot test on the VQA v2.0 dataset, BLIP-2 (ViT-g FlanT5 XXL configuration) achieved an accuracy of 65.0%, significantly higher than the 56.3% of the Flamingo80B model at the same time, proving that the strategy of freezing the pre-trained model parameters can improve efficiency while ensuring performance.

The GIT model proposed by Wang et al. is centered on generative pre-training and transforms the VQA task into an image-driven text generation problem [5]. This model abandons the traditional multi-task branch design and adopts a unified Transformer architecture to process image feature encoding and text answer generation, simplifying the model structure. In the fine-tuning test on the VQA v2.0 dataset, GIT2 (with 5.1B parameters) achieved an accuracy rate of 81.92%, which was 3.11 percentage points higher than the basic version GIT (with 0.7B parameters), indicating the adaptability of the generative architecture to the VQA task and providing an effective path for reasoning on complex problems.

3 Voice VQA

Voice VQA is a task type that integrates voice input on the basis of traditional VQA. It requires first converting the voice questions into a manageable feature representation, and then generating the answers by combining with image information. The technical challenge lies in the noise robustness of the voice signal and the cross-modal temporal alignment [6]. This task provides technical possibilities for hands-free scenarios (such as driver assistance).

In the field of Spoken VQA (Speech-Visual Question Answering), researchers have mainly explored two technical paths. One is the two-stage method based on "first recognition, then answering". The representative model in this field is MUTAN, which adopts the process of "speech recognition and VQA reasoning". It first converts the speech question into text through the ASR engine, and then uses its multimodal tensor fusion (Multimodal Tucker Fusion) [7] mechanism to integrate visual and textual information and generate the answer. The accuracy rate of the answers in a clean speech environment is 58.2%, and it is suitable for static and noise-free scenarios. However, its performance is prone to be affected by the errors in the front-end speech recognition.

To enhance the noise robustness of the model and simplify the processing procedure, another technical approach is the end-to-end model. One representative of this is the MAVEx model proposed by Chen et al. [8]. This model achieves the direct fusion of speech and image features. It directly feeds the speech features extracted by Wav2Vec 2.0 and the image features extracted by ViT into the cross-modal attention module, completely eliminating the intermediate step of speech-to-text conversion. According to its report in the paper "MAVEx: Multi-Modal Answer Validation for Knowledge-Based VQA", this end-to-end model achieves an accuracy of 52.3% in a moderate noise environment, demonstrating that directly processing speech features can effectively enhance the model's robustness in complex environments and is more suitable for voice interaction in mobile scenarios.

4 Image-to-speech QA

Image-to-speech QA takes images (along with optional text descriptions) as input and generates speech-based answers. The core technical links include image understanding, answer generation, and text-to-speech (TTS) synthesis. The key requirements are to ensure the semantic accuracy of the answers and the naturalness of the speech [9]. This task plays an irreplaceable role in scenarios such as visual impairment assistance and smart speakers.

When building an image-to-speech question-answering system, researchers have designed different model architectures based on the requirements of different application scenarios. In scenarios that aim for high audio quality and high accuracy, the representative approach is to combine a powerful visual question-answering model with a high-quality speech synthesis model. For example, using models like FoundationTTS as the speech synthesis backend, its core advantage lies in leveraging generative language models to enhance the naturalness and expressiveness of text-to-

speech conversion. According to its description in the paper "FoundationTTS: Text-to-Speech for ASR Customization with Generative Language Model", when such high-quality TTS models are combined with a powerful VQA model (such as BLIP-2), the overall system achieves a question-answering accuracy of 76.8% on the COCO-QA dataset, and the average opinion score (MOS) of the synthesized speech reaches 3.8, which can meet the requirements of applications with high audio quality.

5 Comparative Analysis and Discussion

5.1 Introduction to Basic Information

The core information of commonly used datasets and evaluation metrics in the MMQA field is as follows, providing a benchmark for task optimization and model comparison:

Core datasets:

VQA v2.0: A benchmark dataset in the field of visual question answering, containing 200,000 images and 1.1 million text questions, covering various daily scenarios and diverse question types, which can comprehensively test the visual-language alignment ability of the model [3].

In response to the specific requirements of voice-visual question answering, the SpokenVQA dataset was born. This dataset contains 48,000 voice questions and is meticulously labeled with key attributes such as accent and noise intensity, specifically designed to support the testing of the model's noise robustness in different acoustic environments, providing a targeted evaluation platform for the optimization of voice VQA models.

COCO-QA: A question-answering dataset constructed based on COCO images, containing 120,000 image-question-answer triples, focusing on object recognition and scene description, suitable for evaluating the semantic accuracy of image-to-speech QA [10].

In terms of evaluation metrics, accuracy is the core indicator for measuring the correctness of answers in VQA and speech-based VQA tasks. It quantifies the model performance by calculating the proportion of samples that match the standard answers, providing a unified criterion for comparing the performance across different models. For speech-based VQA tasks, word error rate (WER) is specifically used to evaluate the accuracy of the speech recognition process. It directly reflects the model's processing precision of the speech signal by calculating the proportion of inserted, deleted, and replaced incorrect words [11].

In the evaluation of the image-to-speech QA system, the speech naturalness score (MOS) assesses the quality of the speech output through manual scoring. This metric employs a 5-point scoring system, with higher scores indicating that the synthesized speech is closer to the natural expression of a human, providing a subjective evaluation standard for the speech synthesis quality of the system.

5.2 Comparison of Tasks and Models

The performance comparison of the main methods in the MMQA field on the core dataset and their applicability scenarios is shown in Table 1, which can directly reflect the advantages and positioning of different technical routes:

Table 1. Comparison of Core Methods for Multimodal Question Answering.

Task	Model	Core Method	Key Performance	Applicable Scene
Visual Question Answering	BLIP-2	Frozen encoder and lightweight bridging module	Zero-shot accuracy: 65.0% (VQA v2.0)	Precise query of image content
Visual Question Answering	GIT2	Unified Generative Transformer Architecture	Fine-tuned accuracy rate: 81.92% (VQA v2.0)	Complex reasoning-based questions and answers
Voice VQA	MAVEx	Voice - Image Feature Direct Fusion	Moderate noise accuracy rate: 52.3% (SpokenVQA)	Mobile scene voice interaction
Voice VQA	MUTAN	Speech Recognition and VQA reasoning	Clean environment accuracy rate: 58.2% (SpokenVQA)	Static noise-free scene
Image-to-speech QA	VQA-FastSpeech	Lightweight Feature Extraction and Fast TTS	Delay: 1.2 seconds, Accuracy: 72.1% (COCO-QA)	Edge device assistance system
Image-to-speech QA	FoundationTTS	Multi-module semantic - prosodic Feature Association	MOS 3.8, accuracy rate 76.8% (COCO-QA)	High-quality audio demand scenarios

5.3 Limitations and Future Prospects

Existing Limitations. Although the current MMQA technology performs well in standard scenarios, there are still significant bottlenecks when it comes to real and complex environments: Semantic alignment accuracy is insufficient: In complex scenarios such as multi-object interaction and abstract problems, semantic mapping between modalities is prone to errors, leading to incorrect answer reasoning. For

instance, in the VQA task, the accuracy rate of judging "the causal relationship of objects in the image" is only 60% of that in the basic scenario [4]. Low noise robustness: In strong noise environments (such as on the street, in factories), the accuracy of speech VQA drops sharply, with WER rising to over 35%, failing to meet the stable usage requirements for mobile scenarios [7]. Lack of naturalness in speech: The output speech from image-to-speech QA often has problems such as harsh rhythm and lack of emotion. The MOS score is generally below 4.0, and there is a gap between the user experience and human interaction [10]. Model deployment is challenging: High-performance models (such as GIT2) have parameter quantities exceeding 5B, and edge devices (such as embedded terminals) are difficult to handle real-time inference, resulting in high deployment costs [11].

Future Research Directions. In response to these bottlenecks, by taking into account the technological trends in the field, we can make breakthroughs in the following directions in the future: Enhancing cross-modal reasoning: Introducing knowledge graphs to assist semantic alignment, and using external knowledge to fill the information gaps between modalities, such as integrating "object attributes - scene associations" knowledge to improve the accuracy of complex problem reasoning [6]. Optimize noise processing mechanism: By integrating speech enhancement techniques (such as noise reduction networks) with an end-to-end architecture, reduce the interference of noise on speech features, and enhance the robustness of speech VQA in complex environments [8]. Improving the quality of speech generation: Integrating an emotion analysis module, so that the output of TTS matches the emotional tone of the problem scenario (such as "urgent issues" corresponding to urgent and urgent speech, "describing issues" corresponding to calm speech), the goal is to increase the MOS score to above 4.2 [10]. Constructing a lightweight unified framework: By employing model compression techniques (such as pruning and quantization) and multi-task sharing strategies, a single model can be adapted to multiple MMQA tasks. While ensuring an accuracy loss of less than 5%, the parameter size is controlled within 1B, thereby reducing the deployment cost on edge devices [11].

6 Conclusions

This study conducts a systematic analysis centered on the core task of multimodal question answering, focusing on three major directions: visual question answering, speech VQA, and image-to-speech QA. It reviews the technological evolution path from traditional two-stage methods to modern pre-training models. By comparing the architecture design and performance data of representative models such as BLIP-2 and GIT, it clarifies the advantages and applicable scenarios of technical routes such as generative architectures and end-to-end fusion - generative models are more suitable for complex reasoning tasks, while end-to-end models are more appropriate for noise-sensitive scenarios.

Research has found that current MMQA technology has achieved high accuracy rates on standard datasets, but there are still significant bottlenecks in semantic

alignment, noise robustness, and deployment lightweighting. These issues limit the application of the technology in complex real-world scenarios and need to be addressed through directions such as knowledge enhancement, noise processing optimization, and model compression.

The value of this study lies in the establishment of a technical comparison framework for the MMQA task, providing a reference for researchers to choose their method paths. At the same time, through the analysis of limitations, it indicates the future innovation directions. With the further development of cross-modal fusion technology, MMQA is expected to achieve more transformative applications in areas such as accessibility services and intelligent interaction, such as providing "real-time scene question answering and natural voice feedback" throughout the entire chain of assistance for visually impaired users.

Reference

1. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443 (2021)
2. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. arXiv preprint arXiv:1801.06146 (2018)
3. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6325–6334 (2017)
4. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12544–12554 (2023)
5. Wang, J., et al.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)
6. Srivastava, A., Memon, A.: Toward robust evaluation: A comprehensive taxonomy of datasets and metrics for open domain question answering in the era of large language models. IEEE Access 12, 117483–117503 (2024)
7. Gao, R., et al.: Adchat-TVQA: Innovative application of LLMs-based text-visual question answering method in advertising legal compliance review. In: 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 176–180. IEEE, Hangzhou, China (2024
8. Babu, N.J., Rajamohana, S.P.: Advancements in knowledge-based visual question answering using large language models: A review. In: 6th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1227–1231. IEEE, Coimbatore, India (2025)
9. Shabtay, N., et al.: Spoken question answering for visual queries. arXiv preprint arXiv:2505.23308 (2025)
10. Ding, D., Yao, T., Luo, R., Sun, X.: Visual question answering in robotic surgery: A comprehensive review. IEEE Access 13, 9473–9484 (2025)
11. Deng, J.: Effective retrieval augmentation for knowledge-based vision question answering. In: 11th International Conference on Behavioural and Social Computing (BESC), pp. 1–5. IEEE, Harbin, China (2024)