Intelligent Question Answering System Based on Multimodal Fusion

Mengcong Zhang

College of Engineering, Environment and Science, Coventry University, Coventry, England, United Kingdom
zhaom18@uni.coventry.ac.uk

Abstract. This article explores the application of multimodal learning in intelligent question answering, emphasizing the importance of integrating data from multiple modalities to fully understand complex scenarios. The article reviews the development of intelligent question answering systems, from the early Turing test to modern models such as BERT and GPT-3, and discusses in detail key multimodal learning technologies, such as data fusion, shared representation learning, and collaborative learning. Using case studies such as Large Language and Vision Assistant, Visual Language Environment, and JBoltAI, the article demonstrates specific applications of multimodal learning in intelligent question answering systems. Finally, the article outlines future development directions, including intelligent agentization, personalization, and neural-symbolic fusion reasoning. These directions will promote the application of multimodal learning technology in more fields, providing users with smarter and more convenient services. Multimodal learning overcomes the information bottleneck of a single modality, enabling machines to possess 'synaesthesia' capabilities closer to humans, enabling explainable decisions in high-trust scenarios such as healthcare, finance, and justice. It also lowers the barrier to entry for AI services and promotes industrial transformation.

Keywords: Multimodal Learning, Image-text Intelligent Question Answering, Data Fusion, Shared Representation Learning, Collaborative Learning.

1 Introduction

The digital age has made the fusion processing of image and text information one of the key research directions in the field of artificial intelligence. Image-text intelligent question answering requires the model to be able to understand and process image and text information at the same time and accurately answer questions related to the image content. Multimodal learning has shown great potential in the task of visual question answering (VQA). Regarding 'multimodality', this article mainly adopts the definition of 'modality' by O'Halloran of the National University of Singapore [1]. Compared with the division of multimedia data such as images, voice, and text, 'modality' is a more fine-grained concept. Different modalities can exist under the same medium. However, traditional single-modality methods have difficulty in effectively fusing

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_36

image and text information, resulting in poor question answering results. Multimodal learning can more comprehensively understand complex scenes by integrating multiple modal data, providing a new approach to solving this problem. In recent years, multimodal learning methods based on pre-trained models have made significant progress. For example, the CLIP model [2] achieves alignment between images and text through contrastive learning, providing powerful feature representation for image-text intelligent question answering tasks. Furthermore, the Transformer-based multimodal model can effectively capture the complex relationships between images and text, further improving question-answering performance. In practical applications, multimodal learning has broad application value in image-text intelligent question-answering, such as in smart education and intelligent customer service, enabling users to provide smarter and more convenient services.

This article primarily introduces some of the basic theories of multimodal learning, reviews the development history of intelligent question-answering systems, and briefly summarizes multimodal fusion, shared representation learning, and collaborative learning techniques within multimodal learning. Regarding the application of image-text multimodal learning in intelligent question-answering systems, the LLaVA model, the VLE environment, and the JBoltAI framework are selected as representative examples for their overview. Finally, this article provides a basic overview of current cutting-edge research directions and future development trends in image-text multimodal learning intelligent question-answering systems. This article summarizes some theoretical techniques for multimodal learning, as well as case studies of the application of image-text multimodal learning in intelligent question-answering systems. This paper hopes to provide inspiration and assistance for future research.

2 Multimodal Learning and Intelligent Question Answering System

2.1 Overview of Multimodal Theory

Human cognition is multimodal. When individuals perceive a scene, they often rapidly absorb visual, auditory, and even olfactory and tactile signals, then fuse and process them to achieve semantic understanding. Multimodal machine learning methods are more closely aligned with how humans perceive the world. Generally speaking, multimodality can take three forms: First is multimedia data describing the same object. This includes, for example, videos, images, audio, and text describing a specific object in the internet environment. Second is the same type of media data from different sensors. This includes, for example, image data generated by different medical imaging devices and data about the same object detected by different sensors in the Internet of Things. Third is ideographic symbols and information with different data structures and representations. These include structured and unstructured data units describing the same object; formulas, logical symbols, function graphs, and explanatory text describing the same mathematical concept; and word vectors, bags of words, knowledge graphs, and other semantic symbolic units describing the same meaning.

2.2 Development History of Intelligent Question Answering Systems

The origins of intelligent question-answering systems can be traced back to the 1950s, when the Turing test asked the question 'Can machines think?' [3]. In the 1960s, MIT's ELIZA used pattern matching and script rules to simulate conversations between psychologists, proving that 'pseudo-understanding' can also create the illusion of interaction [4]. At the same time, systems such as BASEBALL [5] and LUNAR [6] encapsulated database queries into natural language interfaces, laying the foundation for the 'dedicated question-answering' paradigm. In the 1970s and 1980s, knowledge engineering emerged. Stanford University's MYCIN and DRAGON used manually constructed 'if-then' rule chains to conduct medical diagnosis and equipment maintenance Q&A, demonstrating 'explainable reasoning' for the first time, but were limited by the bottleneck of expert acquisition [7]. With the birth of the Internet in the 1990s, FAQFinder [8] and START [9] began to use large-scale text as a source of answers, introducing TF-IDF [10] and shallow grammatical analysis to achieve 'retrieval-based question answering', but they could only return ready-made sentences and lacked generation and logical capabilities. In 1999, the TREC QA track was established to provide standard evaluation for the system and promote joint research between the information retrieval and computational linguistics communities [11].

After 2000, statistical machine learning matured. Researchers used maximum entropy and CRF to classify questions and recognize named entities [12]. They used dependency syntax to find answer sentences and proposed the 'answer ranking' framework for the first time [13]. In 2007, IBM Watson integrated millions of lines of knowledge base and more than 100 algorithms to defeat the human champion in 'Jeopardy!' [14]. This proved that large-scale parallel retrieval plus machine learning can solve open domain fact-based problems, marking the victory of 'system-level integration'. Around 2010, open knowledge graphs such as Freebase [15] and DBpedia [16] appeared. The system mapped questions to SPARQL queries, opened the 'semantic parsing' route, and could answer multi-hop facts, but was limited by the graph coverage and structural completeness.

In 2013, Word2Vec brought distributed representation. For the first time, the question-answering model put the question and the candidate answer into the vector space for similarity matching [17], alleviating the vocabulary gap. In 2015, the attention mechanism [18] and the Seq2Seq framework made 'generative answering' possible. Datasets such as MS MARCO [19] encouraged the model to synthesize natural paragraphs rather than extract the original text. In 2018, BERT refreshed all reading comprehension rankings with bidirectional encoding. The pre-training plus fine-tuning paradigm quickly replaced the traditional pipeline [20]. The system no longer relied on independent word segmentation, syntax, reference and other modules. End-to-end gradient learning dominated the market. In 2019-2020, the number of parameters of language models such as T5 [21] and GPT-3 [22] jumped to the hundreds of billions. Prompt engineering made "zero-shot question answering" a reality. Private domain knowledge can be temporarily injected into the context prompt without retraining.

Starting in 2021, large models will be integrated with external retrieval to form a 'retrieval-augmented generation (RAG)' architecture [23]: first, relevant paragraphs are

recalled in real time using dense vector indexes, and then the large model generates comprehensive answers, taking into account timeliness, credibility, and professional compliance. At the same stage, products such as ChatGPT and Wenxin Yiyan introduced multi-round dialogue management, plug-in tools, and reinforcement learning human feedback (RLHF) [24], expanding question-answering from single-round facts to 'autonomous agents' that can execute code, call APIs, and draw charts. In 2023-2024, multimodal question answering [25] will become a new frontier. The system can simultaneously understand text, tables, images, and videos, and realize 'model as a service' in high-value scenarios such as medicine, law, and finance. At the same time, knowledge editing [26], controllable generation, and privacy computing technologies are used to solve the problems of hallucination, copyright, and data security, and promote the evolution of question-answering systems towards credibility, explainability, and controllability.

3 Techniques and Methods of Multimodal Learning

3.1 Multimodal Fusion Technology

Early fusion methods primarily involve data-level fusion, which is performed before the data is input into the model. This involves combining several data sets into a single data set through various means, and then feeding the data into the model. The advantages of this method are its simple model structure and complexity, the ability to fully leverage complementary modal information, and the ability to perform inference in a single forward pass. However, its disadvantages include dimensionality explosion, alignment difficulties, strong modality bias, high computational complexity, and poor interpretability.

Mid-term fusion methods involve feature-level fusion. The model first extracts features from the data, and then fuses these extracted features. Specifically, vectors extracted from text, images, and audio using CNNs, Transformers, and other methods are first aligned in time and space, undergoing dimensionality reduction. These vectors are then integrated into a unified representation using operators such as tensor fusion, cross-modal multi-head attention, or graph attention, and then trained end-to-end.

The late fusion method is decision-level fusion, also known as output-level fusion, which allows each modality to first train a classifier independently, and then independently produce a prediction vector at the end of each modality path. Then, the probability is integrated at the output layer using strategies such as voting, weighted averaging, Bayesian rule, or logistic regression. The typical approach is to use a lightweight model (ResNet for image processing, BERT for text processing, and MFCC-CNN for audio processing) for each modality to generate logits, which are then dynamically weighted and output by a layer of the learnable gated network. For example, in the early warning of sepsis in the ICU, the vital sign waveform model, the laboratory text model, and the image CNN model each give the risk probability [27]. After weighted average fusion, the modality interpretability is maintained while the redundant information when the modality is missing is utilized, achieving a higher AUC and clinical deployability than a single modality.

3.2 Shared Representation Learning Technology

Shared representation learning maps different modalities, such as word embeddings of text, convolutional features of images, Mel-spectrogram frames of audio, and spatiotemporal blocks of video, to the same high-dimensional manifold by constructing a unified semantic vector space. Cross-modal Transformers (such as LXMERT [28], CLIP [2], and VideoCoCa [29]) are often used in implementation. Local features are first extracted using a modality-specific encoder, and then aligned and interacted through a shared multi-head cross-attention layer, supplemented by contrastive loss or mask reconstruction loss constraints, so that the text description, image, and corresponding sound of a certain target are close to each other in space. A typical example is MedCLIP [30] in the medical field, which shared encoding of X-rays and radiology reports. The zero-shot chest X-ray diagnosis AUC reached 0.85, an 8% improvement over the single-modality ResNet [31]. It can also complete the diagnosis only through the image when the report is missing, reflecting the advantages of cross-modal retrieval, zero-shot reasoning, and missing modality robustness brought by the shared space.

3.3 Collaborative Learning Technology

The core of collaborative learning is to allow different modalities to serve as 'teachers' and 'students' to each other during training. Through a unified optimization objective, they continuously exchange gradients or pseudo-labels, thereby acquiring complementary cross-modal knowledge while maintaining their own unique network structures. Specifically, a 'modality-anchor-positive-negative' framework can be constructed based on contrastive learning: for a video, synchronized speech is used as the text anchor, visual frames as positive samples, and any frames or text from different videos as negative samples. InfoNCE loss is used to maximize the mutual information between the anchor and the positive sample and minimize the similarity with the negative sample. Bidirectional distillation can also be used, where the more powerful modality (such as a high-resolution image CNN) outputs soft labels to the weaker modality (such as a low-sampled audio network). In turn, the gradient of the weaker modality helps the image network focus on areas related to audio semantics.

4 Application Cases of Multimodal Learning in Intelligent Question Answering Systems

4.1 Large Language and Vision Assistant

Large Language and Vision Assistant (LLaVA) [32] is a large multimodal model that connects a visual encoder and a language model for general vision and language understanding. It uses CLIP ViT-L/14 as a visual encoder and Vicuna as a language decoder, and aligns visual features with the language model through a trainable linear projection layer. LLaVA adopts a two-stage training strategy, first performing feature alignment pre-training and then end-to-end fine-tuning. Its innovation lies in the first application of instruction tuning to multimodality, and also proposes the LLaVA-Bench

benchmark. LLaVA performs well in tasks such as visual question answering, image description, and visual reasoning, and can be applied to intelligent question answering systems, visual dialogue, image editing, video understanding and other fields.

4.2 Visual Language Environment

The Visual Language Environment (VLE) is a visual language environment designed to enhance the understanding and generation capabilities of artificial intelligence models by fusing visual information and language interaction. It typically combines advanced visual encoders with powerful language models, enabling models to process visual inputs such as images and videos and generate relevant natural language descriptions, answer questions, or execute visual commands. The core advantage of the VLE lies in its multimodal fusion capabilities, which enable it to understand complex visual scenes and accurately express and interact through language. This environment performs well in tasks such as visual question answering, image description, and video understanding, providing a more natural and efficient way for human-computer communication in intelligent interactive systems. It is widely used in smart assistants, autonomous driving, education, and other fields.

4.3 JBoltAI

JBoltAI is a Java-based artificial intelligence framework designed specifically for enterprise business scenarios, providing powerful intelligent interaction and automation capabilities. It supports function invocation, seamlessly integrating AI with existing enterprise systems to automate business processes. For example, in e-commerce customer service scenarios, AI can access order management systems to obtain real-time logistics information and provide feedback to users. JBoltAI also features event chain orchestration. Through visual node design, enterprises can build complex workflows and achieve full process automation. It also has extensive applications in education, supply chain, human resources, and other fields, such as smart admissions consulting assistants and intelligent tour guide assistants. JBoltAI's multimodal processing capabilities and RAG technology enable it to process multiple data formats and integrate with enterprise-specific knowledge bases for precise decision-making.

5 Frontiers and Future Directions

5.1 Agent-based and Interactive Systems

The next generation of image and text models will close the loop between seeing, thinking and acting, autonomously clicking, capturing or querying web pages, robots and AR scenes to harvest paired data online. Through real-time interactive self-supervision, they collect high-quality paired data, transforming training from offline crawling to online exploration. Reinforcement learning plus causal interventions let them test hypotheses, align symbols with pixels and maintain an explanatory memory that users can interrogate in natural language. This intelligent agentization promises to overcome the three bottlenecks of data, robustness, and interpretability, upgrading large image and text models from 'answer machines' to 'collaborative digital partners'.

5.2 Personalized and Context-Aware

Future image-text models will enable personalized contextual awareness tailored to each user. Through federated learning or local caching, private signals such as user photo albums, chat logs, location, and time will be fed into a lightweight adaptation layer, adjusting image-text retrieval and generation preferences in real time. The model integrates on-device sensors to perceive light, scene, and emotion, dynamically switching between visual encoding styles and textual tone, and providing traceable explanations that are user-interpretable and revocable. Balancing privacy and utility, personalized multimodal assistants will become the core interface for portable AR glasses, in-car HUDs, and home robots.

5.3 Neuro-Symbolic Reasoning

Image-text models are evolving from embedding matching to a hybrid 'neural-symbolic' reasoning approach: neural networks rapidly extract open semantics, while symbolic engines perform interpretable logical operations on external knowledge graphs. The two are mutually trained through differentiable or reinforcement learning cycles. For complex queries such as 'attributes of objects not directly in the image' or 'chronological order of events across images', the system first invokes the visual module to generate candidate symbols, then verifies consistency using logical rules, feeding the results back to the neural module to fine-tune the representation. This approach balances robust perception with rigorous reasoning, promising verifiable and error-correctable multimodal AI, paving the way for high-risk scenarios like scientific computing and legal diagnosis.

6 Conclusion

Multimodal learning breaks through the information bottleneck of a single modality, enabling machines to possess a 'synaesthesia' capability closer to that of humans. In terms of performance, unified representations deliver significant accuracy gains and demonstrate strong generalization capabilities in zero- and few-shot scenarios. In terms of security and trustworthiness, conflict detection and uncertainty estimation enable explainable decision-making in high-trust scenarios such as healthcare, finance, and justice. In terms of accessibility and cost-effectiveness, a lightweight collaborative framework can run in real time on mobile phones and automotive chipsets, lowering the barrier to entry for AI services. Furthermore, multimodal learning is crucial for industrial transformation: the combination of AIGC and embodied intelligence is reshaping trillion-dollar markets such as content creation, education, retail, and manufacturing.

In the future, multimodal learning technology will continue to expand in depth in areas such as trustworthiness and explainability, efficient architectures, interference resistance and robustness, continuous and incremental learning, cognitive-driven learning, and human-machine collaboration. In terms of trustworthiness and explainability, a unified framework for cross-modal causal reasoning and uncertainty management will be constructed to address black-box decision-making and algorithmic

bias. In terms of efficient architecture, by exploring the automated lightweight network of NAS+MoE and the Transformer-CNN hybrid structure, a 10x compression ratio and 1/10 energy consumption are achieved. When facing incomplete and noisy scenarios, robust contrastive learning, adaptive fusion, and test-time adaptation technologies are developed for missing modalities, weak alignment, and label noise. Regarding continuous and incremental learning, lifelong multimodal learning can be supported by studying efficient parameter updates of large models, knowledge distillation, and memory replay mechanisms. In terms of cognitive drive and human-computer collaboration, by introducing cognitive science priors, human feedback reinforcement learning (RLHF) and a multimodal interactive closed loop are realized to create the next generation of intelligent agents that are dialogue-capable, correctable, and co-creative.

References

1. O'Halloran, K.L. (ed.): Multimodal Discourse Analysis: Systemic-Functional Perspectives. Bloomsbury Publishing, London (2006)
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
3. Turing, A.M.: Computing Machinery and Intelligence. Mind LIX(236), 433–460 (1950)
4. Weizenbaum, J.: ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. Commun. ACM 9(1), 36–45 (1966)
5. Green, B.F. Jr., Wolf, A.K., Chomsky, C., Laughery, K.: Baseball: An Automatic Question-Answerer. In: Papers Presented at the May 9–11, 1961, Western Joint IRE-AIEE-ACM Computer Conference, pp. 219–224 (1961)
6. Woods, W.A.: Progress in Natural Language Understanding: An Application to Lunar Geology. In: Proc. of the June 4–8, 1973, National Computer Conference and Exposition, pp. 441–450 (1973)
7. Shortliffe, E. (ed.): Computer-Based Medical Consultations: MYCIN, vol. 2. Elsevier, New York (2012)
8. Hammond, K., Burke, R., Martin, C., Lytinen, S.: FAQ Finder: A Case-Based Approach to Knowledge Navigation. In: Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, pp. 69–73 (1995)
9. Katz, B.: From Sentence Processing to Information Access on the World Wide Web. In: AAAI Spring Symposium on Natural Language Processing for the World Wide Web, vol. 1, pp. 997–1000. Stanford University, Stanford (1997)
10. Sparck Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. J. Doc. 28(1), 11–21 (1972)
11. Voorhees, E.M.: The TREC-8 Question Answering Track Report. In: TREC, vol. 99, pp. 77–82 (1999)
12. Li, X., Roth, D.: Learning Question Classifiers. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)
13. Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G.: Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In: Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–47 (2003)

14. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., et al.: Building Watson: An Overview of the DeepQA Project. AI Mag. 31(3), 59–79 (2010)
15. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: Proc. ACM SIGMOD International Conference on Management of Data, pp. 1247–1250 (2008)
16. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: International Semantic Web Conference, pp. 722–735. Springer, Berlin, Heidelberg (2007)
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013)
18. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473 (2014)
19. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A Human-Generated Machine Reading Comprehension Dataset (2016)
20. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. NAACL-HLT 2019, pp. 4171–4186 (2019)
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. 21(140), 1–67 (2020)
22. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165 (2020)
23. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474 (2020)
24. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al.: Training Language Models to Follow Instructions with Human Feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744 (2022)
25. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Kalyan, A.: Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. Adv. Neural Inf. Process. Syst. 35, 2507–2521 (2022)
26. Mitchell, E., Lin, C., Bosselut, A., Manning, C.D., Finn, C.: Memory-Based Model Editing at Scale. In: International Conference on Machine Learning, pp. 15817–15831. PMLR (2022)
27. Alam, M.U., Rahmani, R.: FedSepsis: A Federated Multi-Modal Deep Learning-Based Internet of Medical Things Application for Early Detection of Sepsis from Electronic Health Records Using Raspberry Pi and Jetson Nano Devices. Sensors 23(2), 970 (2023)
28. Tan, H., Bansal, M.: LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv preprint arXiv:1908.07490 (2019)
29. Yan, S., Zhu, T., Wang, Z., Cao, Y., Zhang, M., Ghosh, S., et al.: Video-CoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners. arXiv preprint arXiv:2212.04979 (2022)
30. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In: Proc. EMNLP 2022, pp. 3876–3886 (2022)
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. Adv. Neural Inf. Process. Syst. 36, 34892–34916 (2023)