ISSN 2616-7433 Vol. 6, Issue 12: 63-68, DOI: 10.25236/FSST.2024.061211

A Review of the Development of Multimodal Large Models

Zheng Miaomiao[1], Gao Yi[1,2,3,*]

[1]Xizang University for Nationalities, Shaanxi, Xianyang, 712082, China
[2]The Collaborative Research Center for Language and Writing Education in Ethnic Regions, Shaanxi, Xianyang, 712082, China
[3]Xizang Key Laboratory of Optical Information Processing and Visualization Technology, Shaanxi, Xianyang, 712082, China
**Corresponding author*

Abstract: *With the continuous advancement of deep learning technology, multimodal large language models built on large-scale language models and large-scale vision models have been making breakthroughs and achieving significant accomplishments in the field of natural language processing. The concept of general artificial intelligence and the explosive popularity of ChatGPT have brought large language models into people's daily lives. These models are typically based on the Transformer architecture, enabling them to handle and generate large amounts of text data while demonstrating strong language understanding and generation capabilities. As multimodal large models progressively enhance their language understanding and reasoning abilities, the application of instruction fine-tuning, context learning, and chain-of-thought tools has become increasingly widespread. This paper mainly analyzes the key technologies and development trends of multimodal large models, as well as the numerous challenges they face.*

Keywords: *Deep learning; Multimodal large models; Transformer; challenges*

1. Introduction

In today's rapidly evolving landscape of artificial intelligence technology, multimodal large models have become a significant force driving technological innovation. Currently, multimodal large model technology is in a stage of rapid development and transformation, with native multimodal large models emerging as new research hotspots. These models aim to break the limitations of processing information from a single modality, achieving deep integration and understanding of various modalities of data. This not only propels the development of artificial intelligence technology towards a more natural and human-like intelligence direction but also enables widespread applications in multiple fields such as autonomous driving, medical diagnosis, educational domains, and entertainment content generation, thereby injecting new vitality into socio-economic development. At the same time, the challenges posed by multimodal large models are becoming increasingly prominent, such as issues with data quality and annotation, cybersecurity risks, hallucination problems, and privacy protection.

Multimodal large models can accept one or multiple types of data inputs and produce a variety of outputs that are not limited to the data types of the input algorithms. The rapid proliferation of multimodal large models is striking due to their significant enhancement of human-computer interaction and their ability to mimic everyday human communication. With the swift adoption and acceptance by a broad user base, many large technology companies, startups, and governments are investing and competing to steer the development of generative artificial intelligence[1].

1.1 Research Background and Significance

Multimodal large models have propelled the development of technology, and with the enhancement of computing power and the surge in data volume, machine learning techniques, especially deep learning, have seen rapid development[1]. Traditional single-modal models often fail to provide sufficient contextual information when faced with complex real-world tasks. The emergence of multimodal large models effectively addresses problems that single-modal models cannot solve. Multimodal large models differ from previous artificial intelligence and machine learning [2,3]. Although

ISSN 2616-7433 Vol. 6, Issue 12: 63-68, DOI: 10.25236/FSST.2024.061211

artificial intelligence has been widely applied to people's lives, the output of most algorithms neither requires nor involves user participation; these platforms attract attention by curating user-generated content. Another distinction between multimodal large models and other types of artificial intelligence lies in their versatility. Previous and existing artificial intelligence models are mostly designed for specific tasks, thus lacking flexibility. Multimodal large models, trained on various datasets, can be used for multiple tasks, offering higher accuracy and robustness[4,5].The research on multimodal large models has profound significance at both the technical and application levels. It not only enhances the performance of models in complex tasks but also promotes the emergence of intelligent interactions and new applications, while also driving the integration of multiple disciplines.

1.2 Domestic and International Research Status

Since the release of ChatGPT [5], Large Language Models (LLMs) have come into the limelight across various industries and have garnered widespread attention. Both internationally and domestically, the research and development of multimodal large models have achieved significant milestones. Globally, tech giants like Google and Facebook have spearheaded the development in this field, while in China, companies such as Alibaba and Baidu have invested substantial resources into the research of multimodal models, successively launching a series of influential models, such as Tongyi Qianwen, Tongyi Wanxiang, and Wenxin Yiyan, among others. These models have demonstrated significant advantages in cross-modal fusion capabilities, parameter volume, and performance, providing robust technical support for multiple industries.

2. Multimodal Large Model Fundamental Theory

2.1 Concept and Characteristics of Multimodal Data

Modality, as a means of describing or perceiving the diversity of things, covers all aspects of information sources and forms of representation. Whether it is the natural sensory ways such as touch, hearing, vision, and smell through which humans perceive the world, or the media and sensors in various technical devices that rely on information transmission, their methods of information acquisition or presentation can be considered as an independent modality. Compared to single-modality models, multimodal large models have several distinct characteristics: stronger comprehensive understanding ability, cross-modal generation and retrieval, enhanced contextual understanding ability, and improved task adaptability.

2.2 Key Technologies of Multimodal Large Models

In the in-depth exploration of the field of artificial intelligence today, research on multimodal large models has become a key force in pushing the boundaries of technology. The focus of this field is deeply concentrated and systematically covers several core aspects, including data collection for pre-training, construction of base models, self-supervised learning and model optimization training[6], fine-tuning for downstream tasks and transfer learning, parallel computing for large models[7], and acceleration of inference.

2.2.1 Pre-training Data Collection

When constructing large-scale models, the quality and richness of pre-training data directly relate to the potential for model performance improvement. Since multimodal large models require far more diverse and high-quality data than single-modal models, the difficulty of acquisition increases significantly. Therefore, exploring low-cost and efficient strategies to mine and construct cross-modal aligned datasets has become an urgent priority.

2.2.2 Base Model Construction

The performance enhancement of large models also faces bottlenecks. Therefore, designing more computationally efficient network architectures, and even exploring alternatives to the Transformer (the structure of large language models based on the Transformer is shown in Figure 1), has become the key to breakthroughs. The advantage of large models lies not only in their strong fitting ability to large datasets but also in their capacity to capture tacit knowledge. However, establishing a clear connection between this tacit knowledge and human common sense remains to be achieved. Thus, the research focus is on developing new models that can effectively integrate implicit and explicit knowledge to

ISSN 2616-7433 Vol. 6, Issue 12: 63-68, DOI: 10.25236/FSST.2024.061211

enhance predictive accuracy and efficiency.

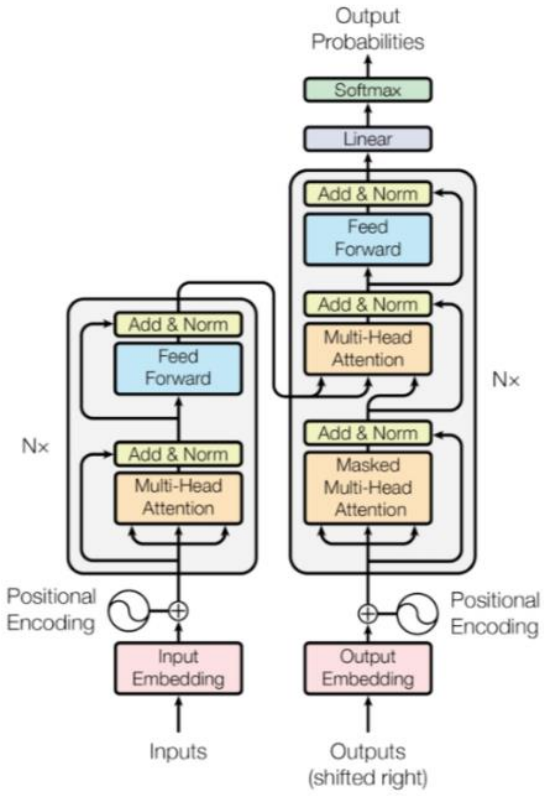

Figure 1 The structure of large language models based on the Transformer

2.2.3 Self-supervised Learning and Model Optimization Training

For large models, optimizing training strategies is equally important as refining model structures. Traditional end-to-end training methods based on regression or contrastive loss are effective, but they lack the trial-and-error and feedback mechanisms present in the human learning process. Therefore, exploring the integration of reinforcement learning into the self-supervised learning process, using environmental feedback to guide model training, has become a new research hotspot.

2.2.4 Downstream Task Fine-tuning and Transfer Learning

The application of pre-trained models in specific domains often requires fine-tuning to adapt to particular tasks. This process leverages a small number of samples to awaken the vast knowledge accumulated during the pre-training phase. Therefore, developing efficient fine-tuning strategies is crucial for fully leveraging the performance of large models. The principle of the fine-tuning step is illustrated in Figure 2.

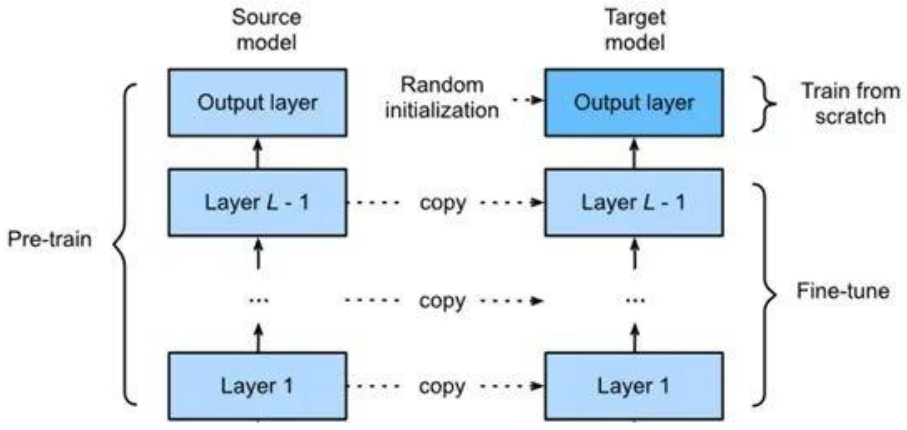

Figure 2 The principle of the fine-tuning step

ISSN 2616-7433 Vol. 6, Issue 12: 63-68, DOI: 10.25236/FSST.2024.061211

2.2.5 Large Model Parallel Computing and Inference Acceleration

In addition to effectively learning from data, large models must also be able to learn quickly. It is necessary to design distributed parallel training methods tailored for ultra-large-scale models.

2.3 Basic Principles of Multimodal Large Models

2.3.1 Multimodal Data Representation

The primary task of multimodal large models is to learn how to effectively represent data from different modalities. This includes extracting features from each modality and representing these features as vectors or higher-level representations. For instance, when processing image data, convolutional neural networks can be utilized to efficiently extract image features; for textual data, advanced models such as recurrent neural networks or Transformers can be employed; and for audio data, technologies like automatic speech recognition can be used for feature extraction.

2.3.2 Multimodal Data Fusion

Multimodal information fusion technology is a method of integrating information from different modalities, aimed at enhancing the decision-making confidence and performance of machine learning and artificial intelligence systems. After extracting features from different modalities, multimodal large models need to fuse these features into a unified representation for subsequent learning and reasoning.

3. Application Fields of Multimodal Large Models

3.1 Applications in Education

Multimodal large models can provide customized learning resources and pathways based on students' study habits, interests, and abilities, achieving personalized teaching. In response to the urgent needs in the field of education, we first construct a general large model for the education field. Subsequently, through a meticulous downstream task adaptation process, we further refine this general large model into three multimodal educational large models, each with its own distinctive features. These three types of multimodal educational large models focus on three core areas: automatic generation of teaching resources, support for human-computer collaborative processes, and intelligent assistance for teacher instruction, forming three typical application scenarios within the education field.

3.2 Applications in Healthcare

On the basis of "big data + high computing power + strong algorithms," by flexibly applying electronic health records, laboratory test results, and medical texts, and other multi-source information, tailored smart medical service solutions can be provided for diverse healthcare scenarios and specific task requirements, ensuring both efficiency and personalization.

3.3 Applications in Intelligent Manufacturing

In the manufacturing industry, multimodal large models can use image and video recognition technologies to conduct real-time defect detection on products in the production line, thereby enhancing product quality and production efficiency.

3.4 Applications in Remote Sensing and Geographic Information Fields

Multimodal large models can classify and interpret remote sensing data such as satellite images and aerial photography, providing comprehensive geographic information and environmental monitoring capabilities. By integrating multimodal information like urban imagery and traffic data, the models can assist urban planners and managers in decision analysis, optimizing urban layout and traffic management.

ISSN 2616-7433 Vol. 6, Issue 12: 63-68, DOI: 10.25236/FSST.2024.061211

4. Challenges Faced by Multimodal Large Models

4.1 Data Quality and Annotation Issues

The annotation issues of multimodal large models, such as increased difficulty in annotation, inconsistent annotation standards, high annotation costs, and difficulty in ensuring annotation quality, are equally complex and challenging.

4.2 Cybersecurity Risks

As people become increasingly reliant on artificial intelligence, these technologies may become targets for malicious attacks and hacking activities, where some systems could be shut down, and training data could be manipulated to alter their performance and responses.

4.3 Hallucinations

Current high-performance language models such as ChatGPT and GPT-4, despite demonstrating exceptional natural language processing capabilities, must confront a serious challenge—the issue of hallucination, which manifests as the frequent output of text containing factual errors or ambiguous content. This phenomenon poses a significant threat to the reliability of model applications in fields that are knowledge-intensive.

4.4 Privacy Protection

When people use multimodal large models, they may not pay attention to privacy. Users who utilize multimodal large models for other purposes often share sensitive information. This shared data may not necessarily disappear quickly, as companies might leverage the data that users share on multimodal large models to improve their artificial intelligence models. The issue of multimodal large models sharing information with other users of the same model is related, whether it is because other users explicitly request the multimodal large model to disclose such information, or the multimodal large model mistakenly discloses someone else's chat logs. Therefore, if an individual's identifiable personal information is input into a multimodal large model, there is a risk that it could be leaked to third parties.

5. Conclusion

The technology of large models has ushered in the era of general artificial intelligence, which is of epoch-making significance and will redefine the information society. The development of multimodal large models demonstrates the immense potential of artificial intelligence technology. From early task-specific single-modal models to the current general multimodal models, although there are still pressing issues to be resolved in large model technology, the continuous advancement of related technologies has brought us more application scenarios and possibilities. In the future, with the popularization of multimodal pre-training and further technological development, multimodal large models will play a significant role in more domains, propelling the further development of artificial intelligence technology.

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C] // Proc of the 30th Annual Conf on Neural Information Processing Systems. New York: Curran Associates, 2017: 5990−6008.
[2] Brown T B, Mann B, Ryder N, et al. Language Models are Fewshot Learners[OL]. arXiv Preprint, arXiv:2005.14165.
[3] Chen Y C, Li L, Yu L, et al. Uniter: Universal Image-text Representation Learning[OL]. arXiv Preprint, arXiv:1909.11740.
[4] Zhang Z, Zhang A, Li M, et al. Automatic Chain of Thought Prompting in Large Language Models[OL]. arXiv Preprint, arXiv: 2210.03493.
[5] Peter J.WORTH.Word Embeddings and Semantic Spaces in Natural Language Processing[J]. International Journal of Intelligence Science,2023,13(1):1-21.
[6] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[DB/OL].https://cdn.openai.

ISSN 2616-7433 Vol. 6, Issue 12: 63-68, DOI: 10.25236/FSST.2024.061211

com/research-covers/language-unsupervised/language_understanding_ paper.pdf, 2018.
[7] Ouyang Long, Wu J, Jiang Xu, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022. 35, 27730−27744.