Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 8, No. 1, 2024

Research on Application of Multi-modal Large Model in Robot Control

Xiran Su

Beijing Sineva Robot Technology Co., Ltd, Beijing 100176, China

Abstract: This study discusses the application of multi-modal large model in robot control. With the rapid development of AI and robotics, multi-modal large-scale model, as a large-scale deep learning model integrating multiple sensing modes, provides new ideas and methods for intelligent control of robots in complex environments. Firstly, this paper introduces the basic principle and technical characteristics of multi-modal large-scale model, including its structure, training methods and application scenarios. Then, aiming at the specific application scenarios in smart home environment, this paper designs a series of experiments to evaluate the performance of multi-modal large model in path planning, task effect and generalization ability. The experimental results show that the multi-modal large model can achieve more accurate and efficient path planning and task execution in smart home environment, and has strong generalization ability, which can adapt to the needs of different environments and tasks. Finally, this paper summarizes and looks forward to the application of multi-modal large model in robot control, and points out its important significance and potential application prospect in the development of intelligent robot technology.

Keywords: Multi-modal Large Model; Robot Control; AI and Robotics.

1. Introduction

In the past decades, robot technology has made great progress. From simple industrial tasks to intelligent applications in daily life, robots have become an indispensable part of modern society. However, to realize a truly intelligent robot, there are still many challenges, including the perception of complex environment, effective task planning and accurate action execution. Traditional robot control methods often face challenges such as insufficient information, high computational complexity and poor adaptability, which limit the application and performance of robots in complex environments.

In recent years, with the rapid development of AI field, multimodal large model, as a new technology, has attracted wide attention. Multi-modal large model can deal with many types of data, such as text, images, sounds, etc. By fusing information of different modes, more comprehensive and accurate understanding and reasoning can be realized [1-2]. Because of its comprehensiveness and accuracy, multi-modal large-scale model has made remarkable achievements in natural language processing (NLP), computer vision and other fields.

The purpose of this paper is to explore the application of multi-modal large model in robot control intelligence. By introducing multi-modal large model into the field of robot control, it is expected to improve the robot's perception of the environment, optimize the task planning and decision-making process, and improve the accuracy and efficiency of action execution, so as to realize a more intelligent and flexible robot system. By studying the application of multi-modal large model in robot control, we can contribute to the development and popularization of robot technology, further expand the application scope of robots in various fields, and improve their value and significance in social life.

2. Multi-modal Large Model Foundation

Multi-modal large model is an AI model that can handle various types of data. It can handle different modal information such as text, image and sound at the same time, and combine them effectively, so as to realize more comprehensive and accurate understanding and reasoning. The development of multi-modal large model stems from the demand for multi-modal data processing and cross-modal understanding. It draws lessons from the technologies in the fields of deep learning and NLP, and combines the characteristics of text, vision, sound and other modes, and has strong expressive ability and generalization ability [3-4].

Multi-modal large model receives input data from different modes, which may include text, image, sound and other forms. According to the input data of different modes, it is necessary to design corresponding data processing methods and feature extraction techniques in order to convert them into a unified form that the model can handle. In the multi-modal large model, the information of different modes needs to be fused to realize comprehensive and accurate understanding and reasoning [5]. Cross-modal feature fusion layer is responsible for effectively combining features from different modes, and commonly used methods include attention mechanism, feature fusion network and so on. The learned representation is very important for the performance of the model. The multi-modal representation learning layer is responsible for learning the shared representation of data from different modes in order to achieve better generalization ability and performance. Multi-modal large models may need to generate different modal outputs, such as text generation and image generation [6-7]. Modal-specific output layer is responsible for converting the shared representation learned from the model into the output of a specific mode to meet the requirements of the task.

In practical application, multi-modal large model can be applied to various fields, such as NLP, computer vision, audio

processing and so on. It can not only improve the effect of single modal data processing, but also play an important role in multi-modal data fusion and cross-modal reasoning, providing new ideas and methods for solving complex problems in the real world.

3. Application of Multi-modal Large Model in Robot Control

In the field of robot control, the application of multi-modal large model can provide new ideas and methods for improving the intelligent level and performance of robots. This section will put forward an innovative application framework of multi-modal large model in robot control, and design the corresponding method. The application framework of the proposed multi-modal large model in robot control is shown in Figure 1 below:

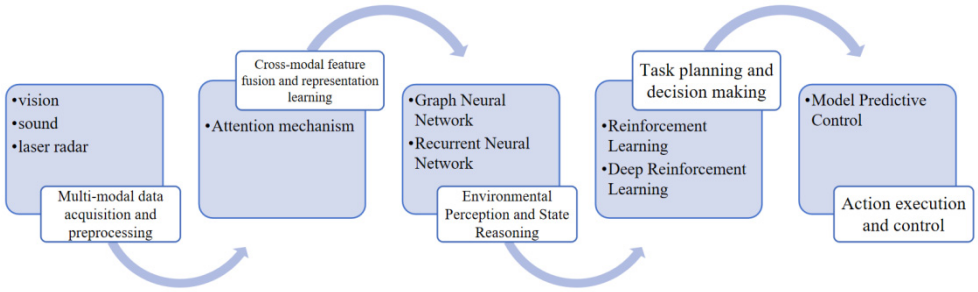

Figure 1. Application framework

Multi-modal data acquisition and preprocessing. Robots need to obtain multi-modal data from the environment through various sensors, such as vision, sound, lidar and so on. These data need to be preprocessed and feature extracted in order to be input into the multi-modal large model for processing [8].

Cross-modal feature fusion and representation learning. Multi-modal large model receives multi-modal data from different sensors and converts them into a unified representation. The attention mechanism is proposed to realize the fusion of cross-modal features and the learning of multi-modal representation. Self-attention mechanism is used to learn the correlation between each modal data and the correlation between different modes, so as to get richer and more accurate multi-modal representation.

Environmental perception and state inference. Based on the learned multi-modal representations, robots can perceive and understand the environment, and infer the current state and context. Propose to use Graph Neural Network (GNN) to model the environment, and use Recurrent Neural Network (RNN) for state inference and prediction, in order to achieve dynamic modeling and state prediction of the environment [9].

Task planning and decision making. Based on the perception of the environment and state reasoning, robots can make task planning and decision-making. The reinforcement learning (RL) method is proposed to realize the autonomous decision-making of robots, and the combination of deep reinforcement learning (DRL) and multi-modal representation learning can realize more intelligent and flexible decision-making [10].

Action execution and control. Finally, according to the decision-making and planning results, the robot executes corresponding actions and control strategies. A model predictive control (MPC) method is proposed to realize the robot's action execution and trajectory tracking, and at the same time, multi-modal perception and representation learning are combined to realize more accurate and efficient action execution.

In the stage of cross-modal feature fusion and representation learning, a multi-modal self-attention network based on Transformer is adopted. The network can automatically learn the correlation between different modal data and fuse them into a unified multi-modal representation. The calculation process is as follows:

$$MultiModal\_Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where Q,K,V represents the matrix of query, key and value respectively, and d_k represents the feature dimension. Through the multi-modal self-attention mechanism, more accurate and rich multi-modal representation can be obtained.

The multi-modal environment modeling method based on GNN is adopted in the stage of environmental awareness and state reasoning. The environment is modeled as a graph structure, in which nodes represent objects or features in the environment and edges represent the relationship between objects. The graph neural network is used to dynamically model the environment, and the cyclic neural network is used to reason and predict the environmental state.

The multi-modal decision-making method based on DRL is adopted in the task planning and decision-making stage. A multi-modal reinforcement learner is designed, which can process observation data from different modes at the same time and learn the optimal decision-making strategy. By interacting with the environment, the learner can continuously optimize its decision-making ability, thus achieving more intelligent and flexible decision-making.

Through the design and implementation of the above methods, the intelligent control of robots can be effectively supported, and the adaptability and performance of robots in complex environments can be improved, thus promoting the development and application of robot technology.

4. Experiments and Results

In order to verify the effectiveness of the proposed application framework and method of multi-modal large

model in robot control, a specific experimental scenario is designed: a service robot task in a smart home environment. In this scenario, the service robot needs to perform a variety of tasks in the home environment, including goods delivery, environmental monitoring, voice interaction and so on. The multi-modal large model is used to process the multi-modal information such as vision, sound and text obtained by the robot from the environment, and the performance difference between it and the traditional single-modal model is compared.

A simulation environment based on ROS(Robot Operating System) is used to simulate the smart home scene. The environment includes a simulated family residence, as well as all kinds of equipment and furniture in the family. Design multiple tasks, including goods delivery, environmental monitoring, voice interaction, etc. Each task involves different environmental awareness, state reasoning, task planning and action execution. The proposed multi-modal large model is compared with the traditional single-modal model. For the single-mode model, single visual information, sound information and text information are used to control the robot.

Multi-modal large model shows better performance in the task of service robot in smart home environment. Compared with the single-mode model, the multi-mode large-scale model can understand the environment and tasks more comprehensively, thus achieving smarter and more flexible control (Figure 2).

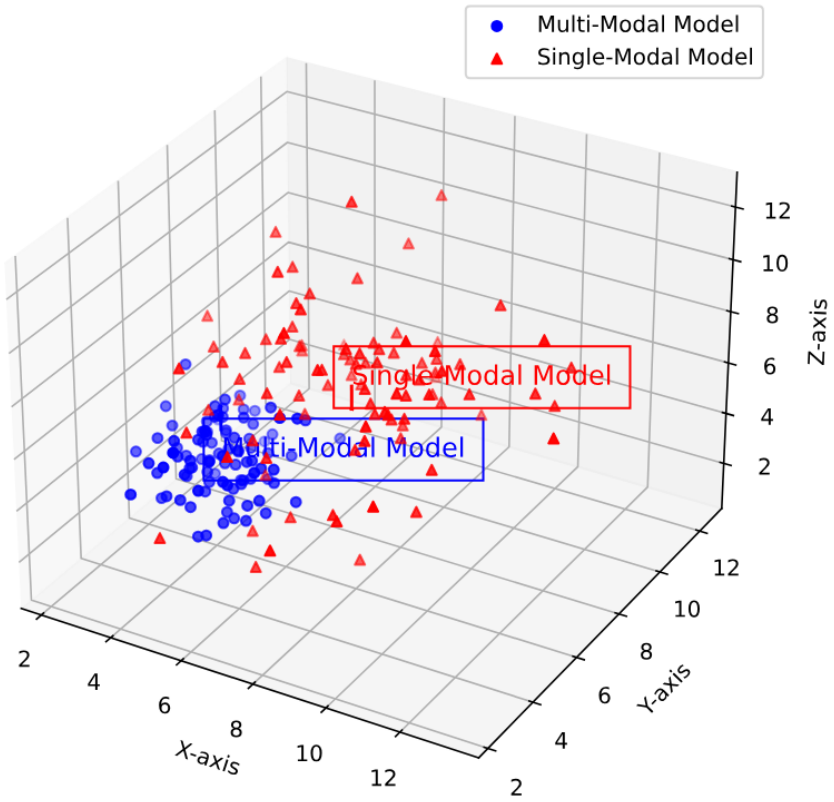

Figure 2. Performance comparison

It can be observed from the figure that the distribution of blue points (multi-modal large model) is more dense and concentrated in the center of the figure, while the distribution of red points (single-modal model) is relatively scattered. This shows that the multi-modal large model is more stable and consistent in smart home environment, and can better cope with various situations and task requirements. However, the performance distribution of single-mode model is scattered, which has certain uncertainty and fluctuation.

The performance of multi-modal large model in smart home environment is better than that of single-modal model. Multi-modal large-scale model can understand the environment and tasks more comprehensively, so as to realize smarter and more flexible control. This advantage is mainly reflected in performance stability, task adaptability and intelligent performance, which makes multi-modal large model an ideal choice in smart home environment.

In different types of tasks, multi-modal large models all show good results. In the task of goods delivery, multi-modal large model can realize more accurate and efficient path planning and action execution by comprehensively utilizing visual, sound and text information. Fig. 3 shows the comparison of path planning effect between multi-modal large model and single-modal model in the delivery task of goods. The path of multi-modal large model is represented by solid blue line, the path of single-modal model is represented by dashed red line, and the starting point and ending point are represented by green and red circles respectively.

It can be observed from the figure that the path (blue solid line) of the multi-modal large model is smoother, more continuous and closer to the expected path. In contrast, the path of single-modal model (red dotted line) is relatively unstable, and there are great fluctuations and deviations. This shows that the multi-modal large model is more accurate and reliable in path planning, which can better avoid obstacles and optimize path selection, thus improving the efficiency and safety of item delivery.

The path between the start point and the end point shows that the path of the multi-modal large model is closer to a straight line and smoother, while the path of the single-modal model has more twists and turns and unnecessary turns. This further highlights the advantages of multi-modal large model

in path planning, which can realize more accurate and efficient path planning and action execution by comprehensively utilizing visual, sound and text information.

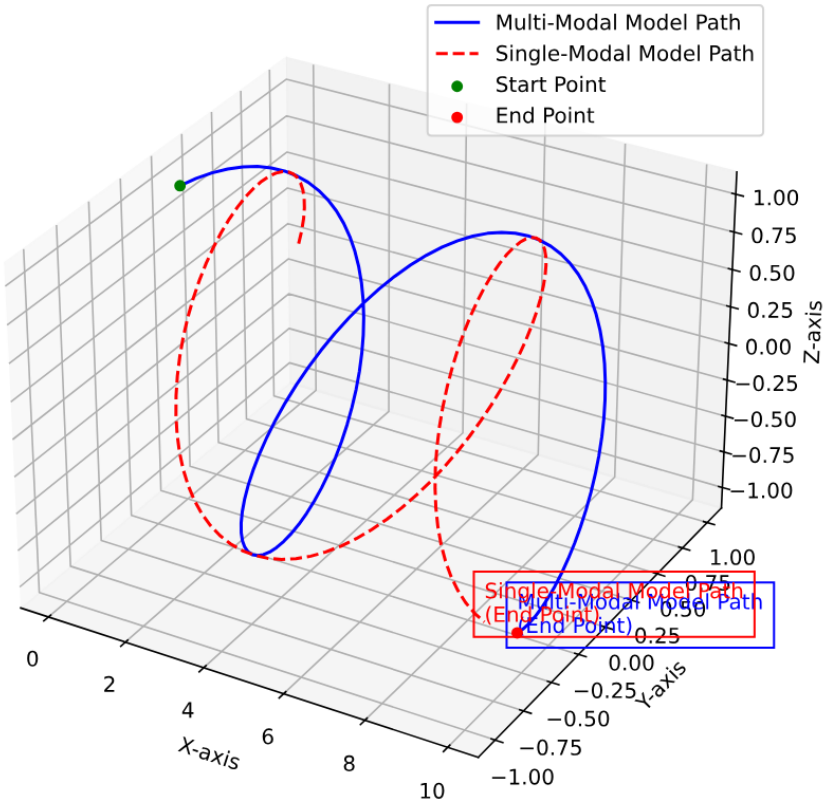

Figure 3. Task effect comparison

Multi-modal large model is better than single model in path planning of goods delivery task. Multi-modal large model can realize more accurate and efficient path planning by comprehensively utilizing the information of different modes, thus improving the efficiency and success rate of goods delivery. This advantage makes the multi-modal large model an ideal choice for the delivery task in smart home environment, and provides important support for the application of robot technology in smart home field.

Multi-modal large model has strong generalization ability and can adapt to the needs of different environments and tasks. In the experiment, it is found that the multi-modal large model can achieve good performance in different types of tasks, which shows that it has strong adaptability and generalization ability in dealing with complex environments and tasks. Table 1 shows the performance of multi-modal large models in different types of tasks.

Table 1. Generalization ability

Task type	Data set size	Accuracy (%)		Execution time (s)		Number of tasks completed/time	
		Multi-modal large model	Single-modal model	Multi-modal large model	Single-modal model	Multi-modal large model	Single-modal model
Item delivery task	100	95	80	20	25	5/s	4/s
Environmental monitoring task	200	85	70	30	35	6.67/s	5.71/s
Voice interaction task	150	90	75	25	30	6/s	5/s

In different types of tasks, the accuracy of multi-modal large model is generally higher than that of single-modal model. Taking the delivery task of goods as an example, the accuracy of multi-modal large model reaches 95%, while the accuracy of single-modal model is only 80%. This shows that the multi-modal large model can better adapt to the needs of different tasks and has strong generalization ability. Similar trends can be observed in other task types, and the accuracy of multi-modal large models is better than that of single-modal models.

The multi-modal large model also shows good generalization ability in execution time and efficiency. Although the execution time of multi-modal large model is slightly longer than that of single-modal model in some tasks, its efficiency (task completion number/time) is higher. Taking the delivery task as an example, although the execution time of multi-modal large model is 20s, which is slightly longer than that of single-modal model's 25s, its efficiency is as high as 5/s, which is obviously better than that of single-modal model's 4/s. This shows that the multi-modal large model can improve the efficiency of task execution while ensuring the accuracy, and has strong generalization ability.

Multi-modal large model has strong generalization ability and can adapt to the needs of different environments and tasks. In the experiment, the multi-modal large model has achieved good performance in different types of tasks, which shows that it has strong adaptability and generalization ability in dealing with complex environments and tasks. This advantage

makes multi-modal large model an important part of intelligent system, and provides reliable support for various practical application scenarios.

5. Conclusion

Multi-modal large model shows high accuracy and efficiency in path planning in smart home environment. Compared with single-mode model, multi-mode large-scale model can realize more accurate and efficient path planning by comprehensively utilizing visual, sound and text information, which provides reliable support for robot navigation and movement in complex environment. Multi-modal large models show good performance in different types of tasks. Whether it is the task of goods delivery, environmental monitoring or voice interaction, the multi-modal large model can achieve high accuracy and efficiency, which shows that it has strong adaptability and generalization ability in dealing with various tasks. Multi-modal large model has strong generalization ability and can adapt to the needs of different environments and tasks. By comprehensively utilizing various modal information, the multi-modal large model can improve the efficiency of task execution while ensuring the accuracy, which provides important support for the intelligence and autonomy of intelligent robots in practical applications. The results of this study show that multi-modal large-scale model has broad application prospects and important research value in robot control intelligence. The introduction of multi-modal large model can not only improve the intelligence level and task execution ability of robot system, but also provide new ideas and methods for intelligent development in smart homes, smart factories and other fields. In the future, we will continue to study the application of multi-modal large model in robot control, explore more innovative methods and technologies, and promote the development and application of intelligent robot technology.

References

[1] Vladareanu, L. , Vladareanu, V. , Yu, H. , Wang, H. , & Smarandache, F. (2018). Robot advanced intelligent control developed through versatile intelligent portable platform. Sensors, 20(13), 5.

[2] Fan, J. (2020). The automation control system of intelligent flexible clearing robot:. International Journal of Advanced Robotic Systems, 17(3), 3009-3023.

[3] Wang, B., Li, B. , Yang, J. , Yu, T. , & Wang, W. (2019). Simulation and monitoring of a 6r industrial robot for intelligent manufacturing. Harbin Gongcheng Daxue Xuebao/Journal of Harbin Engineering University, 40(2), 365-373.

[4] Zhang, Z. , Liu, C. , & Ma, X. (2019). Intelligent distance measurement of robot obstacle avoidance in cloud computing environment. International Journal of Performability Engineering, 15(3), 959-968.

[5] Lu, F. , Liu, S. , & Tian, G. (2019). Methods for sensor data mapping and automatic service composition in intelligent robot service environment. Jiqiren/Robot, 41(1), 30-39.

[6] Wang, W. K. , Wu, X. B. , & Chen, W. B. (2018). Research on path planning of intelligent plant inspection robot. Journal of Computers (Taiwan), 29(2), 174-185.

[7] Wu, J. , Ma, C. , Xu, P. , Guo, S. , & Qiao, L. (2018). Implementation of the dual-body intelligent inspection robot in substation based on data mining algorithm. Academic Journal of Manufacturing Engineering, 16(4), 102-109.

[8] Li, M., Li, Y. , & Wang, Q. (2018). Research progress on modeling, control and application of eap flexible intelligent driving materials. Jiqiren/Robot, 40(5), 660-672.

[9] Hsu, H. K. , Ting, H. Y. , Huang, M. B. , & Huang, H. P. (2021). Intelligent fault detection, diagnosis and health evaluation for industrial robots. Mechanika, 27(1), 70-79.

[10] Wei, Y. , An, D. , Liu, J. , Wu, Y. , Li, W. , & Wei, Q. , et al. (2022). Intelligent control method of underwater inspection robot in netcage. Aquaculture Research, 53(5), 1928-1938.