HiLL PUBLISHING

Advances in Computer and Communication, 2025, 6(2), 81-86
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Research on Intelligent Decision-making Methods for Complex Scenarios Based on Multimodal Large Models

Lili Song

Shenzhen Nuoyi Fluid Technology Co., LTD., Shenzhen 518107, Guangdong, China.

How to cite this paper: Lili Song. (2025) Research on Intelligent Decision-making Methods for Complex Scenarios Based on Multimodal Large Models. *Advances in Computer and Communication*, 6(2), 81-86. DOI: 10.26855/acc.2025.04.005

Received: March 19, 2025
Accepted: April 16, 2025
Published: May 13, 2025

***Corresponding author:** Lili Song, Shenzhen Nuoyi Fluid Technology Co., LTD., Shenzhen 518107, Guangdong, China.

Abstract

With the gradual development of technology, traditional intelligent decision algorithms face major challenges in a complex environment in which different data from multiple sources, great uncertainty, and dynamics coexist. This paper deals with the powerful capabilities of multimodal large models (MLMs) in the processing of multimodal data such as text, images, audio, etc. First, the basic nature of MLMs and the characteristics of complex scenarios are discussed, and existing MLMs-based decision-making training methods and their limitations regarding data generalization and interpretation are also discussed. At the same time, a new framework is proposed that contains three key components: first, it uses pre-treatment techniques such as noise reduction and intermodal alignment to solve the noise problem when multiple modes of transport are used; second, it combines early and late coupling with attention mechanisms to achieve mixed low-level signal Fusion. trait correlation and high-level semantic interdependence; third, it uses advanced learning optimization strategies to adapt to dynamic environmental feedback. The study used the nuScenes autonomous vehicle dataset and data from the field of Medicine and wellness mimic-3 for evaluation. The accuracy rate of decision-making in autonomous vehicles reached 92.5% and in medical services 89.3%, which is 4-5% higher than in the conventional method of combining data. This study demonstrates the importance of adaptive modal weighing and iterative strategy optimization, as well as that the data privacy model has urgent issues that require a solution, such as improved interpretability and extensibility. This is relevant to study.

Keywords

Multimodal large models; complex scenarios; intelligent decision-making; data fusion; reinforcement learning

1. Introduction

The rapid development of Information Technologies has led to an increase in the volume and types of data in all areas of life. Complex conditions include many factors, modes, and dynamic interactions that are quite common in practical applications. In autonomous driving, vehicles must obtain information from multiple sources, such as cameras, and make immediate decisions in a variety of situations, for example, in dynamic road conditions. In terms of Intelligent Healthcare, doctors need to integrate a multitude of patient data for diagnosis and treatment planning. In this complex scenario, traditional single-mode decision-making methods that use only one data source often prevent the disclosure of hidden rich information and its overarching importance, leading to unsatisfactory decision-making results.

The construction of a multimodal large model is based on deep learning and natural language processing technology. It can process and integrate data from multiple modes. After a deeper study of the high-level semantic representation of multimodal data, the accuracy of content descriptions and patterns of complex scenarios has been significantly improved. With the support of the multimodal large model, researchers have proposed many intelligent decision-making methods for complex scenarios. In some scenarios, the effect is good, but there are still a lot of problems in this area. There are many unknown factors associated with the efficient integration of multimodal data, and improving the efficiency of Model generalization in unknown scenarios is still a challenge. There are still many problems in ensuring the interpretability and fairness of Model decision-making results. problem.

Given these problems, this paper proposes an intelligent decision-making method based on a multimodal large model that can handle complex scenarios. The structure of the article is as follows: Section 2 discusses relevant studies; Section 3 describes the proposed scheme and decision-making algorithm; Section 4 describes the conditions of the experiment and its results; Section 5 discusses difficulties and guidelines for the future; Section 6 provides the entire text.

2. Accompanying work

2.1 Multimodal oversized models

The advent of a new era of digital technology has allowed data to exist in different forms. The study of multimodal large models (MLMs) can be seen as a direct result of extracting useful cooperative information from heterogeneous signals. MLMs using deep learning platforms, including Transformers and hybrid neural networks, are expected to be the power source for the simultaneous processing of multimodal data streams (text, images, sounds, and sensors). The main variables in its architecture are time and method modal coupling, which are usually classified as early coupling, late coupling, and hybrid coupling architectures [1].

The early coupling method integrated multimodal objects at the object level by combining visual-text objects from CNNs and RNNs into a single vector for downstream processing. This method directly integrates bimodal functions, enhancing cross-modal learning. For instance, Visual Genome integrates image regions and embedded text sentences in scene diagram creation, using a large bimodal scale of visual text to define visual language tasks; otherwise, the discriminant of information quantity in multimodal merge functions remains undefined. Due to the association with only a multimodal function, unsynchronized modal features (e.g., images and text) and high noise can result in loss of recognition details.

In the process of late coupling, information about all modes enters independent pipelines (via different network blocks) for processing and is finally combined at the level of decision-making or prediction. Category probability or regression estimate) are integrated by weighted aggregation or voting schemes by a majority of votes or meta models (such as logical regression) to determine how results are generated. This method guarantees the preservation of information about one mode at an early stage of processing in a particular mode, especially when a particular mode is more reliable (for example, tactile sensors in certain specific operating conditions that are more critical than visual data when the robot is operating) and has a significant preference. In medical decision systems, radiological images and texts of patient medical records are usually processed separately using different models, and the results of multimodal prediction are combined using Bayesian networks [2]. The limitation of this method is that a single modal pipeline limits the possibility of deep interaction with multimodal information. In the process of in-depth study of characteristics, although processing in each mode remains independent, different merge points occur before cross-modal transmission and data compression [3]. Modal information is generated gradually through hierarchical learning, without the possibility of early deep integration. This is mainly achieved through independent work, and important cross-modal information cannot be obtained in full. This can lead to the abandonment of the deep interaction of functions in some specific modes, as well as the reduction of opportunities that are important for integrating and improving the cross-modal decision-making/forecasting process.

Although the choice between early and late coupling has attracted attention, the hybrid coupling model has recently been used as a heuristic solution to achieve mode coupling in a hierarchical or iterative manner. Several modes are jointly processed at all levels, rather than aggregation being maintained at the entire level. The tiny coarse-grained elements are generated using independent modal encoders, and attention controls or other splicing Cascades are used for multi-level splicing at a deeper level and ultimately for cooperative processing. An important example is a stand-alone multi-sensor control system. At a lower level, lidar point clouds and RGB images from cameras can be

combined at the same time to know the relative position of the vehicle according to RGB data; at a higher level, point clouds and extended semantic maps are used for cooperative processing and route planning. This method of coevolution and coupling can also be applied to tasks that benefit from multi-level abstraction, such as video subtitles and disaster scene understanding, a semantic multimodal platform for deep learning.

2.2 Intelligent decision-making methods based on multimodal large models

In the new conditions of detailed synthesis of different information flows (mainly based on single-mode input data), multi-mode large models are successfully applied in intelligent decision-making systems in most industries. The application of the recommendation system combines user information (text information about the demographic characteristics of the user), product information (product image characteristics) and interactive information (user text, images, transactions and other behaviors), so that the model's ability to personalize recommendations outweighs simple collaborative filtering methods in a single mode. Alibaba's multimodal recommendation [4] uses transformer-based encoders to integrate single-mode data streams, such as user interaction with product text comments (text), visual information, and transaction history (images), etc., to obtain vector (built-in) views, determining user preferences outside of cold processing. run problems and increase the variety of recommendations to create personalized products that users prefer (higher click rates).

Multimodal models in medicine, including MRI, EHR, and genetic data, aid diagnosis and treatment planning. Paper [5] confirmed multimodal use effectiveness and proposed a late-stage Fusion architecture linking chest X-ray CNN features to longitudinal EMF data. In pneumonia patient classification, the single-mode model's accuracy is 12% higher than normal, showing that reviewing other regimens aids in avoiding bias-induced diagnostic preferences. In autonomous driving [6], a hybrid model combining camera, lidar, and radar sensors has been proposed to aid decision-making and enhance situational awareness. In weak signals, reduce weight, e.g., increasing radar weight in heavy rain.

Although these efforts have allowed progress in combining models to solve complex problems in the real world, there are still limitations. suming that the restrictions are quite rigid, the dynamic characteristics of the everyday life of the scene are significant, and the statistical distribution of various modal data will change over time, such as, for example, the distribution of temperature and traffic accidents varies depending on the season, in different patients may come from different mod or rules). It is not effective in summarizing the ability of invisible data. Static styles or corrective models trained based on past patient data have limitations in dealing with current or future disease variations (these variations are missing from past training data) due to the lack of ability to study the correlation between regimens over time [7]. Most existing Fusion simulation systems take into account uncertainty and ambiguity in dynamic scenarios. There are limitations, such as scarce collision risk sensor data in autonomous driving systems, as well as a lack of value in electronic medical records (EHR), which can lead to the loss of certain potential patient information. Traditional merge models based on deterministic integration of information are unable to describe the uncertainty and ambiguity between different modal data (with fewer labels). So, in the model, too confident, unoptimal, or even wrong decisions are made.

In complex scenarios, agent-agent and agent-environment interactions persist, posing a significant challenge to the generalization ability of existing models due to online dynamics and feedback. City traffic systems must integrate sudden jams and pedestrian patterns to handle behavior and visual disruptions, yet current methods learn in standalone scenarios [8]. Learning scheme based on time-varying distribution, unspecified connectivity, and sequence dependence.

3. An intelligent decision-making system proposed

3.1 Structure review

The intelligent decision-making system presented in Figure 1 is based on a multimodal, large-scale model. Within the platform, three modules were implemented, including the multimodal data preprocessing module, as well as the Model Integration module, and the decision-making optimization module.

3.1.1 Multimodal data pre-processing

At the pre-processing stage, multimodal data must be cleaned, normalized, and aligned. Deleting data can remove noise and foreign values. Normalization of the data makes it possible to bring the data to a single scale, which allows

for an increase in the speed of convergence of the models and their efficiency. Data alignment refers to the synchronization of different modal data in time or space, and this step is necessary for the integration of multimodal information.

3.1.2 Model integration

After pre-processing the data, the second step is to combine a multimodal large model. Within this framework, a hybrid Fusion Strategy is used that combines early and late coupling. First, different submodels are used to obtain the characteristics of each mode, and then a combination of functions on multiple levels is performed. In particular, the original characteristics of different modes are related at the basic level and are used as the same intermediate-level inputs; at the higher level, the attention mechanism relates the main characteristics of different modes, and this mechanism can dynamically assess the importance of the mode.

3.1.3 Optimization of decision making

The last step of the system is to optimize the decision-making process and improve the learning algorithm to optimize the decision-making strategy. Environmental feedback information can update the decision-making strategy and gradually improve the efficiency of decision-making. At the same time, consider the interpretability and objectivity of the decision-making results and add appropriate regular conditions to the target optimization function.

3.2 Algorithms

3.2.1 Algorithm for extracting multimodal characteristics

A deep neural network can derive the characteristics of each mode. For the processing of graphical data, a CNN, text data RNN, or a transformer-based model is used, and the extracted objects are normalized and entered into the Fusion module.

3.2.2 Hybrid coupling algorithm

The hybrid coupling algorithm has two parts: low-level and high-level coupling, with small elements from different modes connected for full coupling layer processing to obtain low-level coupling elements. High-level coupling combines deep functions via concentration, determines attention difficulty, and merges weight with deep functions to form high-level coupling, eventually integrating low-level and high-level coupling into final coupling functions [9].

3.2.3 Algorithm for optimizing the decision-making process

The formulation of the decision-making process is based on deep reinforcement learning. The Markov decision-making process (MDP) is a modeling system in which multimodal functions are integrated into a state, the execution of the decision-making process is an action, and environmental feedback is performed on behalf of the reward. For optimal decision-making strategy training, a deep kindness network (DKN) or strategy gradient method is used to maximize reward accumulation over time.

4. Results of the experiment and analysis

4.1 Experimental setting

To assess the effectiveness of the proposed method, testing was conducted in two complex and representative scenarios: autonomous vehicles and intelligent medical care.

4.1.1 Data sets

Autonomous driving dataset: We used the nuScenes dataset [10], which is a large multimodal autonomous driving dataset containing various sensors, cameras, radars, etc. The content of the marking includes information about the environment, objects, and road signs.

Smart Healthcare dataset: mimic-III dataset is a large-scale clinical dataset that contains patient information from electronic health records (EHR), medical images, and physiological signals. It can be used in Applied Medical tasks such as diagnosing diseases and predicting their effects.

4.1.2 Assessment indicators

Indicators such as accuracy level, readability, recall rate, F1 rating, and speed of operation allow for assessing the effectiveness of the proposed method.

4.1.3 Comparison methods

The proposed method is compared to several existing smart decision-making methods based on multimodal large models. The early fusion model is included in the comparison model, the late Fusion model is also included in the comparison model, and the non-attention mechanism hybrid Fusion model is also compared with itself.

4.2 Results of the experiment

4.2.1 Autonomous driving scenario

Table 1 shows autonomous driving experiment results; the proposed method outperforms existing technology in all criteria with 92.5% accuracy. Early fusion improves accuracy by 5.2%, late fusion by 4.8%, outperforming existing methods in accuracy, sensitivity, and F1 estimation. In terms of computational efficiency, the costs are slightly higher, but the difference between early and late Fusion models remains within range.

4.2.2 Intellectual health care scenario

The experimental results of intelligent health scenarios given in Table 2 are similar to the results of autonomous driving scenarios. The proposed method has the best performance in all assessment indicators. In terms of accuracy, the proposed method can reach 89.3%, which is 4.5% higher than the early fusion model and 4.2% better than the late Fusion model. Accuracy, recall, and F1 rates are also higher than existing methods. In terms of computational efficiency, the proposed method is a little slower, but it can still be applied in real-time applications.

4.3 Analysis of results

The results of the experiments show that the proposed method allows for to increase of the efficiency of intellectual decision-making in complex situations for the following reasons:

Multimodal data coupling: The proposed method uses a hybrid coupling strategy that allows recording the low-level correlation and high-level semantic relationship between multimodal data and provides more complete and accurate information for decision making.

Attention-getting mechanism: The attention-getting mechanism in a high-level Fusion module can target the current scene and dynamically regulate the importance of different modes. Attention is paid to the most effective regimes, which also improve the accuracy of decision-making.

Reinforcement learning optimization: The reinforcement learning method is applied to a decision optimization module that can learn from environmental feedback and adapt to unspecified changes in scene dynamics, thus increasing the efficiency of Model generalization.

5. Discussion

5.1 Problems

Although the proposed method has provided results with promising characteristics (the specific meaning here needs to be supplemented according to the context) in the experimental scenario, there are still plenty of questions to consider in the future.

Privacy: When using decision-making models in applications such as intelligent medicine and finance, data often contains sensitive content. When large-scale multimodal models are applied, the guarantee of data privacy and security becomes especially important. The privacy and security guarantee sets important requirements for the data processing process in a multimodal model.

Model: Multimodal large models are complex and difficult to explain. It becomes quite difficult for users to trust and understand the results of decision-making. Another challenge is to increase the comprehensibility of the model and clarify the decision-making process.

Scalability: The rapid growth of volumes and data types continues to encourage thinking about Model extensibility and the development of scalable algorithms that can efficiently process large-scale multimodal data. This problem has yet to be solved.

5.2 Directions of future research

The suggestive nature of the above problems led the author to think about some future directions of research:

Preserving multimodal learning: exploring privacy protection. Federal training and differentiated privacy can ensure the privacy/security of multimodal data in the implementation of learning and decision-making processes.

Creation: New techniques and receivers, such as attention visualization and rule-setting, can improve the interpretability of multimodal large models. The use of cause-and-effect relationships is also important for ensuring transparency and ease of understanding decision-making processes in large models. At the same time, the requirements for transparency contribute to the visualization of the expression of the internal characteristics of the model. Processing.

Scalable multimodal large models: The development of efficient and scalable architectures and algorithms, distributed computing, and parallel model processing for multimodal large models is particularly important. Processing large multimodal data and complex scenarios also contributes to the complication of the computational process.

6. Conclusion

This work explores intelligent decision-making in complex scenarios using multimodal large models, outlining their key features and introducing a new platform. The platform covers multimodal data preprocessing, model integration, and decision-making optimization, with hybrid Fusion Strategies and advanced learning methods enhancing effectiveness; experiments in two complex scenarios demonstrate superior accuracy and efficiency over existing methods.

Although the proposed methodology has achieved satisfactory results, there are still many challenges and opportunities for further research in this area. We hope that this study will open up a new perspective and technique for intelligent decision-making in complex scenarios, as well as encourage other researchers to delve deeper into this important topic.

References

[1] Guo C. Research on multi-modal trajectory prediction method for intelligent vehicles based on deep inverse reinforcement learning [dissertation]. Shaanxi: Chang'an University; 2023.

[2] Song D. Research on intelligent vehicle anthropomorphic decision-making and planning strategy based on traffic situation cognition [dissertation]. Jilin University; 2024.

[3] Fu W. Research on intelligent decision making of autonomous driving based on multimodal fusion [dissertation]. Heilongjiang: Harbin Institute of Technology; 2023.

[4] Wang T. Research on multi-modal trajectory generation method based on flow model [dissertation]. Liaoning: Dalian Maritime University; 2022.

[5] Xu Y. Research on multi-modal adaptive gait generation method for bipedal robot based on deep reinforcement learning [dissertation]. Shanghai: Shanghai Normal University; 2024.

[6] Wang X. Research on multi-modal fusion 3D multi-target tracking method for intelligent vehicle environment perception [dissertation]. Chongqing: Chongqing University; 2023.

[7] Zhong Z. Research on decision planning method of autonomous vehicles based on dynamic interaction [dissertation]. Guangdong: South China University of Technology; 2023.

[8] Luo F. Research on multi-modal trajectory prediction method based on attention mechanism [dissertation]. Shandong Institute of Business and Technology; 2024.

[9] Cheng H. Research on multi-modal disaster information detection method based on deep learning [dissertation]. Qilu University of Technology; 2023.

[10] Deng W. Research on autonomous navigation system of mobile robot based on multimodal perception fusion in indoor environment [dissertation]. Shanghai: Shanghai Normal University; 2024.