International Journal of Advanced Research in Engineering Technology & Science ISSN: 2349-2819
Impact Factor: 7.10
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Email: editor@ijarets.org Volume-11 Issue-11 November-2024 www.ijarets.org

A REVIEW EXISTING MULTI-MODAL AI SYSTEMS AND IDENTIFY GAPS.

Avinash Alugolu

Research Scholar Sikkim Alpine University

Dr. Prasadu Peddi

Professor, Sikkim Alpine University

Abstract

In recent years, there has been a tremendous increase in the development of multimodal artificial intelligence (AI), which is a set of systems that are capable of integrating and interpreting information that has been gathered from a variety of data sources, including text, pictures, audio, video, and sensor streams. Tasks such as visual question answering, emotion recognition, video captioning, cross-modal retrieval, and human–AI interaction have shown that the current multimodal AI architectures, including transformer-based fusion models, contrastive learning frameworks, and large multimodal models (LMMs), have the ability to perform at a high level. Existing systems continue to encounter persistent restrictions that restrict their robustness, generalizability, and application in real-world situations, in spite of these considerable accomplishments. This research covers the present landscape of multimodal AI systems, encompassing primary model types (early-, late-, and hybrid-fusion models), essential benchmark datasets, and application areas across healthcare, autonomous driving, social media analysis, and education. A number of significant shortcomings have been discovered as a result of the investigation. To begin with, multimodal models frequently have difficulty when the modalities are misaligned, which is when discrepancies across data streams in terms of temporal, spatial, or semantic information lead to a decrease in the accuracy of the model. Secondly, the fact that many systems depend on datasets that are both domain-specific and massive in size makes it difficult to scale them up, and it also makes them problematic to use in situations with limited resources. Third, the present multimodal architectures are still not capable of effectively addressing the missing, incomplete, or noisy modalities that are frequently encountered in actual contexts. Fourth, the interpretability and explainability of multimodal systems are still restricted, which makes it difficult to audit or trust their decision-making processes, particularly in situations where the consequences are severe. Fifth, because of the unevenness of multimodal datasets, present methods frequently demonstrate bias amplification, which results in predictions that are either unjust or incorrect. Finally, multimodal learning has yet to achieve cross-cultural and cross-linguistic generalization, which limits its use in a wide variety of global situations. In spite of the fact that multimodal artificial intelligence has made significant advancements, the domain is still struggling with difficulties pertaining to data alignment, model transparency, cross-modal robustness, bias reduction, real-time performance, and generalization beyond selected benchmarks. It will be necessary to address these gaps in order to build multimodal artificial intelligence (AI) systems of the future generation, which will be more trustworthy, inclusive, and capable of working successfully in open-world situations.

Keywords: *multi-modal, AI, systems, identify*

Introduction

One of the most revolutionary developments in the field of machine learning at the present time is multimodal artificial intelligence (AI), which seeks to combine information from a variety of data sources, including text, images, speech, video, and sensor signals, in order to develop systems that have a more human-like capacity for comprehension. In contrast to unimodal models, which are based on a single kind of input, multimodal systems emulate human cognition by interpreting and mixing a variety of modalities in order to create representations of the world that are both more detailed and more accurate. This paradigm change has been motivated by the substantial advancements that have been made in the field of deep learning, the appearance of massive multimodal datasets, and the advent of foundation models that are capable of processing diverse information streams. There has been a tremendous expansion of the ways that multimodal artificial intelligence is utilized across several fields in recent years. Multimodal models may be utilized in the healthcare industry to aid in diagnosis by merging radiological pictures with electronic health information. Perception systems combine data from a variety of sources, including cameras, LiDAR, and radar, in order to provide safe navigation in the context of autonomous driving. Multimodal artificial intelligence is utilized by social media platforms to identify emotions, disinformation, or dangerous material by employing a combination of visual and linguistic indicators. Multimodal signals, including facial expressions, voice tone, and interaction logs, are applied by educational technology in order to individualize learning experiences for students. The increasing significance and influence on society that multimodal systems has is demonstrated by these advancements. Despite the progress that has been made, the development of successful multimodal artificial intelligence models continues to be an intricate problem. In many cases, data from different modalities not only differ from each other in terms of their structure and dimensionality, but they also vary in terms of their temporal resolution, semantic depth, and noise levels. In order to align and integrate these sources, which are so varied, it is necessary to utilize complex fusion procedures, which range from early fusion and late fusion to attention-based hybrid architectures. In addition to this, the success of multimodal systems is strongly reliant on the availability of big datasets that are of good quality; yet, a number of real-world situations include datasets that are partial, missing, or have poor labels. Multimodal models typically demonstrate variable performance when implemented outside of controlled benchmark conditions as a result of this. The following are a number of recurrent difficulties that have been pointed out by prior studies: modality imbalance, in which one modality exerts a dominant influence on prediction; cross-modal inconsistency, in which modalities are in conflict with one another; restricted interpretability, which makes it difficult to comprehend how various inputs influence decisions; and problems with scalability, particularly in situations with limited resources or low-resource languages. Furthermore, issues of openness, fairness, robustness, and ethical usage of multimodal data have emerged as key considerations in talks on the future of multimodal artificial intelligence. Taking into account the information provided here, it is necessary to conduct a thorough examination of the multimodal AI systems that are now in use in order to comprehend the most advanced techniques, assess important applications, and determine the shortcomings that are preventing the broad adoption of these systems. This introductory section serves as a prelude to a thorough investigation of the currently extant multimodal architectures, learning paradigms, and assessment procedures. Following this research, the next step will be to conduct an analysis of the strengths and weaknesses that characterize the multimodal artificial intelligence landscape as it exists now. The goal of this study is to draw attention to significant deficiencies in the existing research and to provide an overview of possible paths for progress in the development of the next generation of multimodal systems, which are expected to be more resilient, transparent, scalable, and inclusive.

Rationale

Building user interfaces from the ground up with JavaScript, HTML, and CSS guarantees a fluid and easy-to-navigate experience. It enhances responsiveness and aesthetics with the help of jQuery, Popper.js, and Bootstrap. Processing of data, interaction with APIs, and user authentication are all responsibilities of the backend, which is driven by Python and Django. In order to foretell how the market will act in the future, the central machine learning model examines trends in past value using LSTM. Visual Studio Code is the main IDE, and Google Colab is utilized for model help. This platform offers a trustworthy and extensible market research solution by integrating AI-driven predictions with real trading activities. The application of deep learning techniques has enhanced its ability to anticipate financial markets, which is beneficial for both researchers and investors.

Traders, investors, and financial specialists may use historical data to project how much stocks and cryptocurrencies will be worth in the future according to the AI-powered Stock and Crypto Value Prediction System. Due to market instability, traditional value forecasting methods may fail to identify intricate patterns. This system assists users in making data-driven investment decisions and enhances forecast accuracy through the use of machine learning techniques, namely Long Short-Term Memory networks. Django is used for secure authentication and management, and the system allows users to create and manage their profiles. It starts by collecting and preprocessing historical stock and cryptocurrency value data from external APIs in order to analyze time-series data and predict future trends. Visualizations that show both historical and future data, together with interactive charts and graphs, show the predictions. The front end is made easy to use and browse by utilizing jQuery, Bootstrap, JavaScript, CSS, and HTML. In addition, customers may use the system at any time, day or night, via a web-based platform, so they can verify forecasts whenever they choose. Reliable value forecasting is now possible with this system's combination of deep learning and web technologies. It enhances market decision-making and lowers investment risks.

Econometric Models Affect Stock And Crypto Value Prediction System

Analyzing economic events and relationships is done using an econometric model in econometrics. Through the application of statistical methods and economic theory, it quantifies and analyzes relationships among economic variables. In order to comprehend the economy and make well-informed policy decisions, these models are important.

A number of scholars have explored machine learning as a potential replacement for conventional methods of estimating stock values. Preprocessing techniques for financial time series data were the main emphasis of Peng et al. (2019). Various combinations of LSTM model parameters were explored, and data was processed using methods such wavelet de-noising, normalization, and interpolation (Peng et al., 2019). Their results showed that the new model significantly improved prediction accuracy while decreasing computation expenses. In addition, prior studies that attempted to predict stock prices either employed a combination of machine learning and statistical econometric models or utilized many machine learning models all at once. It is common for many models to get better results than a single model. Two model combination approaches, LSTM-RNN and BPA-MLP, were assessed by Achkar et al. (2018). The two most common methods for predicting stock prices are econometric models and algorithms based on machine learning. However, traditional machine learning methods only consider data from a single time

period, overlooking crucial implicit information that varies over time, while econometric methods struggle to handle nonlinear time series problems (Baek and Kim, 2018). Data structured throughout several eras or time series may be amenable to a novel approach called deep learning. Combination models are now all the rage for predicting stock values since they are more accurate than individual models.

Multimodal Representation Learning

In the beginning of multimodal learning research, the goal was to separate characteristics from each modality and then combine them when making a choice. When Ngiam et al. (2011) demonstrated that acquiring shared representations across auditory and visual modalities increases performance on speech-related tasks, it was one of the seminal contributions in the field. Later studies used CNNs and deep autoencoders to extract characteristics that are exclusive to or independent of a certain modality (Srivastava & Salakhutdinov, 2012). Transformative learning frameworks and designs like BLIP, ALIGN, and CLIP (Radford et al., 2021) have been crucial in recent advancements; these systems learn cross-modal connections from massive amounts of image-text pairings. Thanks to their impressive zero-shot and cross-domain generalization capabilities, these models have completely changed the game for multimodal research. In order to address the drawbacks of previous static fusion methods, researchers like Tsai et al. (2019) suggested multimodal transformers that use cross-modal attention processes to dynamically capture interactions across modalities. Despite these advancements, problems such a lack of connections between modalities, misalignment, and an imbalance between modalities are often brought up in the literature. More versatile representation learning algorithms that can deal with missing or noisy modalities are needed, as highlighted in works by Zadeh et al. (2017) and Mai et al. (2021).

Fusion Strategies in Multimodal Systems

Early fusion, late fusion, and hybrid fusion are the three categories into which fusion techniques are usually grouped, and each one has its own set of benefits and downsides. The process of combining modalities at the moment of input is referred to as early fusion, or feature-level fusion. Baltrušaitis and colleagues (2019) shown that while early fusion has the capability to capture low-level interactions, it is also susceptible to excessive dimensionality and noise. The incorporation of modality-specific predictions is a characteristic of late fusion, which makes it resilient to missing modalities; nevertheless, it is less successful in simulating cross-modal interactions. By employing ensemble-based late fusion, Wu et al. (2016) were able to achieve enhanced performance in the categorization of videos. Hybrid fusion, with a particular emphasis on transformer-based cross-attention, has emerged as the most advanced method. The findings of research like Perez-Rua et al. (2020) and Li et al. (2022) demonstrate that hybrid approaches outperform standard fusion procedures, thanks to their utilization of hierarchical multi-level interactions. On the other hand, thorough reviews, such as Zhang et al. (2022), have shown that fusion techniques still struggle with temporal synchronization in tasks including video and audio, semantic inconsistency between text and pictures, and a limited capacity to scale up to big, noisy datasets from uncontrolled situations.

Multimodal Datasets and Benchmarks

Large-scale datasets that are accessible to the public have been the driving force behind the advancement that has been made in the discipline. Among the most important benchmarks are the following: MS-COCO, Visual Genome, VQA, and Flickr30k for vision–language. AVA, LRS3, and Kinetics-Sound are

all examples of audio-visual technology. MOSI, MOSEI, and MELD are examples of multimodal emotion recognition. The following multimodal datasets are used in the healthcare industry: MIMIC-CXR, UK Biobank imaging corpora. Although the use of these datasets has made it possible to carry out systematic evaluation, a number of studies (such as Raji et al., 2020; Escalante et al., 2021) have demonstrated that the datasets that are now in use display demographic imbalances, a lack of cultural variety, and settings that are restricted to certain environments. As a consequence of this, a large number of models do not generalize well when they are applied to actual data.

Application-Specific Multimodal Research

Healthcare: The results of studies conducted by Rajpurkar et al. (2022) and Chen et al. (2023) indicate that the multimodal integration of clinical text and medical imagery enhances both the accuracy of diagnoses and the prediction of prognoses. Nevertheless, there are still some significant limits, such as difficulties with modality that is not present and worries over privacy. Systems that operate without human intervention: Studies like the one conducted by Chen et al. (2017) show that the identification of obstacles is improved when data from LiDAR, radar, and cameras are combined. In spite of this, the research indicates that the fundamental problem is the ability to remain resilient in the face of unfavorable weather conditions and sensor malfunctions. Zadeh and colleagues (2017) demonstrated the significance of modeling interactions by introducing the Tensor Fusion Network (TFN) for the purpose of sentiment and emotion research. According to Hazarika et al. (2020), further study found that memory networks produced improved results; nonetheless, the ability to generalize across cultures remained limited. Multimodal methods have demonstrated their effectiveness in identifying misleading information and damaging content on social media platforms, however research by Gomez et al. (2021) has revealed that these techniques have their limits when the modalities are in conflict with one another or when multimedia posts have been purposefully modified.

UTILIZE OF DEEP LEARNING MODELS FOR MONETARY TIME SERIES FORECAST

Forecasting for financial time series with the goal of estimating future asset prices. There is a plethora of methods to choose from, however the majority of studies have utilized deep learning models in order to forecast the movements of assets. The following factors are included in this study: stock, index, foreign exchange, commodities (including gold and oil), bonds, volatility, and cryptocurrency prices. The primary concepts of these predictions are applicable irrespective of the matter at hand.

The field of financial time series forecasting is home to two distinct ideologies. The first of these is concerned with making accurate predictions about the values in question, while the second ideology is more concerned with making predictions about the direction of movement. Although the ideal use of regression tasks is for the purpose of predicting exact values, a number of attempts to anticipate financial situations prioritize the accurate identification of the direction in which values are moving above the precise prediction of those values. Rather than concentrating on the precise forecasting of values, the primary focus should be on the forecasting of trends, which entails figuring out the direction in which values are anticipated to change. As a consequence of this, we consider trend prediction to be a classification problem. When a certain number of studies solely examine binary outcomes, such as upward or downward movements, but others include a third class, the neutral option, a conundrum involving three classes emerges. In the course of their investigation, Ben Ameur and his colleagues employed a variety of

deep learning models in an effort to make predictions about the prices of commodities for the Bloomberg Index. It was determined that the LSTM models performed more effectively when compared to the other models. Baser and colleagues conducted an investigation that involved the examination of a large number of tree-based models with the objective of determining how effectively these models would be able to predict the price of gold. Among the models that were used were Decision Trees, XGBoost, Gradient Boosting, Random Forest, and AdaBoost. They were able to demonstrate that XGBoost was more accurate than the other models by conducting a study of technical indicators. With the use of statistical and machine learning models, Deepa and her coworkers were able to determine that boosted decision tree regression was the most effective method for estimating the amount of cotton that will be produced in India. Zhao and Yang developed a deep learning system that was capable of making more precise predictions regarding fluctuations in stock prices by incorporating market data and investor sentiment into its calculations. Sentiment analysis and deep learning approaches were both utilized in this strategy.

Recurrent neural networks (RNNs) have the capacity to more effectively boost the grasp of temporal correlations in sequence data than feedforward neural networks due to the fact that they are capable of utilizing previous hidden states in order to assess the inputs that are now being received.

Recurrent neural networks (RNNs) make modifications to hidden states in response to both present and previous inputs at each time step, therefore gradually processing stock value data.

In addition, recurrent neural networks (RNNs) are more effective and more useful for predicting what will happen in the stock market since they have the ability to process inputs of varying lengths and they utilize parameter sharing.

The most appropriate data format for a graph neural network to manage is a graph. They are able to record complicated links and interactions by gathering and updating information from nodes.

By combining their own attributes with those of their neighbors, global neural networks (GNNs) are able to learn graph representations, which in turn enables them to update the states of nodes. When it comes to stock forecasting, there are benefits to using graph neural networks.

Nodes are used to represent the stocks, and edges are used to highlight the connections that exist between them. The study of the characteristics of the nodes and edges that are indicative of stock market movements and interactions has the potential to allow global neural networks (GNNs) to forecast stock prices.

Due to the fact that it offers more precise market data and forecast accuracy than standard time series-based models, this method is particularly suitable for projects that include the prediction of several assets.

Conclusion

The fast development of multimodal AI has opened up new possibilities for creating systems that can interpret and reason using a wide variety of data types, including text, pictures, audio, video, and signals from sensors. This paper emphasizes that multimodal AI is still limited in its potential due to a number of basic restrictions, even if there has been a lot of development in areas such as representation learning, fusion techniques, and domain-specific applications. Visual question answering, emotion identification, cross-modal retrieval, and clinical decision support are just a few of the tasks where current multimodal

architectures, particularly transformer-based and contrastive learning models, have proven to be extraordinary. But these improvements are still not uniformly spread among modalities and use cases. The difficulty of misalignment, which can be either spatial or semantic in nature, is a common subject in the literature and a barrier to effective cross-modal representation acquisition. Similarly, multimodal systems aren't as flexible when it comes to dealing with noisy, inconsistent, or missing input from the real world since they depend on huge, high-quality datasets that are generally domain-specific, which restricts scalability. The management of missing modalities is another important area that needs improvement. This is an issue in many fields, including healthcare, autonomous systems, and platforms for user-generated material. Even while newer models try to fix this by using imputation and modality dropout techniques, they still fail to keep performance stable when important modalities are broken or missing. Training large multimodal models is computationally intensive, which limits their use in low-resource environments and makes real-time deployment challenging. In sensitive applications such as recruiting, medical diagnosis, or monitoring, multimodal datasets often display demographic, cultural, and environmental biases that might compound damages through model projections. Problems with decision auditing and understanding the impact of multiple modalities on model outputs are made worse by the lack of interpretability. A multimodal AI system that is open, equitable, and responsible is desperately needed in light of these worries.

References

1. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.
2. Chen, L., Tai, L., Sun, Q., & Li, B. (2017). Multi-Sensor Fusion for Autonomous Driving Using Deep Neural Networks. IEEE International Conference on Robotics and Automation (ICRA), 1–8.
3. Chen, T., Roy, S., Wang, S., & Wong, A. (2023). Multimodal Learning for Clinical Decision Support Systems: A Review. Journal of Biomedical Informatics, 136, 104258.
4. Escalante, H. J., Martí, L., & Ponce, J. (2021). Bias and Fairness in Multimodal AI Systems: A Critical Review. ACM Computing Surveys, 54(8), 1–35.
5. Gomez, R., Papadopoulos, S., & Kompatsiaris, I. (2021). Multimodal Analysis for Fake News Detection. Information Processing & Management, 58(4), 102553.
6. Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. ACM International Conference on Multimedia, 1122–1131.
7. Li, Y., Zhou, X., & Sun, X. (2022). Hierarchical Cross-Modal Attention for Multimodal Fusion. IEEE Transactions on Multimedia, 24, 2401–2413.
8. Mai, S., Xing, S., & Hu, H. (2021). Modality Dropout for Robust Multimodal Sentiment Analysis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5.
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning (ICML), 689–696.
10. Perez-Rua, J. M., Zhu, X., Zhang, B., & Martinez, B. (2020). Knowing What, Where and When to Look: Efficient Video and Image Understanding via Multimodal Attention. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1–10.

11. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), 8748–8763.
12. Raji, I. D., Smart, A., White, R. N., Mitchell, M., & Gebru, T. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. ACM Conference on Fairness, Accountability, and Transparency (FAT*), 33–44.
13. Rajpurkar, P., Irvin, J., & Lungren, M. P. (2022). Multimodal Medical AI: A Survey and Outlook. Nature Machine Intelligence, 4, 365–378.
14. Srivastava, N., & Salakhutdinov, R. (2012). Multimodal Learning with Deep Boltzmann Machines. Advances in Neural Information Processing Systems (NeurIPS), 2231–2239.
15. Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 6568–6579.
16. Wu, Z., Jiang, Y.-G., Wang, J., Ye, H., & Xue, X. (2016). Multimodal Deep Learning for Video Classification. IEEE Transactions on Multimedia, 18(2), 233–245.
17. Zadeh, A., Chen, M., & Poria, S. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. Conference on Empirical Methods in Natural Language Processing (EMNLP), 1103–1114.
18. Zhang, Z., Cui, L., & Chen, Y. (2022). A Comprehensive Survey on Multimodal Fusion Techniques. ACM Transactions on Multimedia Computing, Communications, and Applications, 18(3), 1–25.