International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 1, Jan-Feb-2024, ISSN (Online): 2395-566X

A Comprehensive Literature Review on Multimodal Large Language Models for Integrated Text, Image, and Speech Understanding

Research Scholar Chintu Kodanda Ramu, Professor Dr.Pankaj Khairnar
Sikkim Alpine University, Kamrang ,Namchi ,Sikkim

Abstract- — AI technology is actually growing very fast and has definitely created big computer programs that can understand difficult written information. Real-world information actually comes in different forms like text, images, and speech, but traditional systems definitely cannot combine these forms effectively. As per this study, we review all research papers regarding Multimodal Large Language Models that can understand text, images, and speech together. This review surely examines how multimodal learning has evolved from old rule-based and machine learning methods to modern deep learning approaches. Moreover, it specifically looks at the shift towards transformer-based architectures in recent years. The study shows that early systems used handcrafted features and could not adapt further, while machine learning methods performed better but were itself limited by manual feature extraction. Deep learning methods like CNNs and RNNs helped machines learn features automatically, but they faced problems in understanding long connections and interactions between different types of data itself. Further research was needed to solve these limitations. Transformer models solved these problems using attention mechanisms, which further led to MLLMs that combine different data types in one framework itself. The review also studies different ways to combine multiple data types, shared spaces for embedding, and cross-modal attention methods as per enhancing better understanding and reasoning abilities. Despite good progress, challenges like data alignment, computational complexity, scalability, and need for large multimodal datasets remain critical problems that require further attention. These issues itself create barriers for better implementation. As per the study findings, there are important research gaps regarding the need for better system designs, improved data combining methods, and practical solutions that can work on a larger scale. Basically, this review gives a complete picture of how MLLMs are developing, what challenges they face, and where they're heading, showing they can bridge the same gap between how humans think and machine intelligence.

Keywords— Multimodal Large Language Models, Multimodal Learning, Text Image Speech Integration, Transformer Models, Cross-Modal Attention, Deep Learning, Representation Learning, Multimodal Fusion, Artificial Intelligence, Contextual Understanding

I. INTRODUCTION

AI has actually changed a lot from simple rule systems to smart models that can definitely do complex thinking tasks. Large language models are actually a major step forward that use transformer designs and definitely work very well at understanding how words connect with each other in text. These models have surely achieved success, but they can only work with text. Moreover, real-world information comes in many different forms like images, sounds, and text together.
Basically, human thinking works by combining different senses like seeing, hearing, and language - it's the same process of putting all these inputs together. Also, when people watch a video, we are seeing that they process the spoken words, visual parts, and background hints all at the same time only. This combined understanding surely helps humans get meaning better than systems that use only one mode. Moreover, it makes the process much more effective. Basically, traditional AI systems cannot do this properly, so they give incomplete understanding of complex data the same way.

We are seeing multimodal artificial intelligence coming up to solve this problem, where different types of data are brought together into one single system only. Multimodal Large Language Models further extend traditional language models by adding image and speech capabilities. This approach itself enables systems to understand and process multiple types of data together. These models actually use attention methods to find connections between different data types, which definitely helps them understand context better and work more effectively. Basically, more multimodal data is available now, so researchers are focusing on building efficient systems that can handle the same type of complex information at scale.Lu et al. [1], Tan and Bansal [2].

II. THEORETICAL BACKGROUND

Multimodal understanding surely follows the basic ideas of representation learning and emotional computing, where computer systems try to understand complex human information. Moreover, these systems aim to process different types of data together to make sense of how humans communicate and express themselves. As per the analysis, each method has different features regarding how information is presented - text follows a sequence with symbols, speech uses time and sound, and images work with space and many dimensions. Combining these different methods into one system is surely a difficult task because they have different structures and ways of showing information. Moreover, these differences make it hard to bring them together effectively.
As per early computer methods, they used manual feature extraction which limited their ability regarding capturing complex patterns. Basically, deep learning made neural networks automatically extract features and learn representations the same way humans do in layers. Convolutional Neural Networks were used for image processing, while Recurrent Neural Networks and Long Short-Term Memory networks were used for sequential data like speech and text itself. Further, these networks helped process different types of data effectively.

However, these architectures had limitations in modeling long-range dependencies and cross-modal interactions itself, which further restricted their performance. Transformer models solved these problems by using self-attention mechanisms that help the system focus on important information in sequences. This approach further improved how models process data by allowing them to understand relationships within the sequence itself. Basically, these mechanisms help model relationships within and between different modalities in the same efficient way.

Multimodal Large Language Models further extend transformer architectures by representing different modalities in a shared embedding space itself. This unified representation further enables cross-modal reasoning, allowing the model itself to understand relationships between text, images, and speech. We are seeing that these abilities are only needed for work that requires putting together understanding and making decisions.Chen et al. [3], Li et al. [4].

III. REVIEW OF PREVIOUS STUDIES

Early and Traditional Approaches
Early studies on multimodal understanding and emotion analysis surely depended on rule-based systems and manually created features. Moreover, these approaches used handcrafted methods to process different types of data together. These methods actually used set rules to understand feelings from text, face expressions, and how people speak. They definitely followed fixed patterns to read emotions. They actually gave basic ideas, but they definitely could not change with new situations or work with real complicated data. These methods surely gave basic understanding, but they could not adapt well or work with difficult real-world information. Moreover, their performance remained limited in practical situations. We are seeing that these methods gave basic understanding only, but they could not change with different situations and handle complex real-world information properly. They actually gave basic ideas, but they definitely could not change with new situations or work with difficult real data. Further, thesedata.Sun et al. [5].

Machine Learning Approaches
Machine learning actually brought data-based models that definitely made performance better by learning patterns from information. Algorithms like Support Vector Machines and Decision Trees were surely used widely for classification tasks. Moreover, these methods became popular choices for organizing data into different categories. These models used features from each modality and gave better accuracy than traditional methods. Further, the approach itself showed improved performance compared to earlier techniques. These methods still needed manual work to create features and further faced problems with complex data itself.

Deep Learning Advancements
Deep learning surely helped computers understand different types of data much better by automatically finding important patterns. Moreover, this technology made it possible for machines to learn features on their own without human help. We are seeing that CNNs take out space features from pictures, while RNNs and LSTMs only work with time patterns in speech and text. As per the research, hybrid designs that combine these models gave better results regarding the integration of spatial and temporal data.

Despite these improvements, deep learning models faced further challenges like high computational needs and limited ability to capture long-range dependencies itself. Also, basically, combining different types of data in the same framework was still complex. Deep learning models had the

same problems like needing too much computing power and difficulty understanding long connections in data. Further, combining different types of data in one system itself remained difficult. These deep learning models still faced problems like high computing needs and poor ability to understand long connections in data. Surely, combining different types of data in one system was still difficult. Moreover, deep learning models had problems like needing too much computing power and could not handle long-distance connections well. We are seeing that bringing together different types of data in only one system was still very difficult.Radford et al. [6], Ramesh et al. [7].

Transition to Multimodal Systems

The problems with single-mode systems surely pushed researchers to create multimodal methods that use many data sources together. Moreover, these new approaches work better by combining different types of information. We are seeing that early systems only used simple methods to combine information from different types of data by joining their features together. Moreover, as per the results, performance got better, but it could not catch the deep connections between different types of data.Jaegle et al. [8], Li et al. [9], Jia et al. [10]

Basically, advanced fusion methods with attention and hierarchical approaches were introduced to solve the same limitations. These methods surely give different levels of importance to each type of data automatically, which helps combine them better. Moreover, this approach leads to improved understanding of the context.

Table 1: Role of Modalities in Multimodal Large Language Models

Modality	Input Type	Representation	Contribution to Understanding	Challenges
Text	Words, Sentences	Token embeddings	Semantic and contextual meaning	Ambiguity, sarcasm
Image	Pixels, Frames	Visual embeddings	Object recognition and scene understanding	Noise, occlusion
Speech	Audio signals	Acoustic features	Tone, speech patterns, context	Background noise
Multimodal Integration	Combined inputs	Shared embedding space	Holistic understanding and reasoning	Alignment, fusion complexity

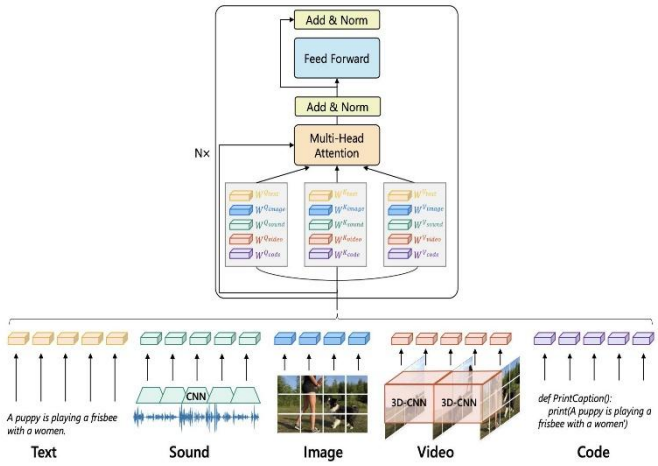

Figure 1: Architecture of Multimodal Large Language Model for Integrated Text, Image, and Speech Understanding

Transformer-Based Models and MLLMs

Transformer models surely changed how we handle different types of data together by making it easier to process sequences and multiple data forms. Moreover, these architectures brought a major shift in multimodal learning approaches. Self-attention mechanisms help models capture long-range dependencies and relationships within modalities and further across different modalities itself.

Further, multimodal Large Language Models surely expand this idea by combining text, image, and speech together. Moreover, they create a single unified system that can work with all these different types of data. Moreover, as per the design, these models convert different types of data into tokens within one shared space, regarding cross-modal reasoning and contextual understanding capabilities. As per recent studies, they have shown good results in tasks regarding multimedia analysis, conversational AI, and combined information processing.

As per requirements, these models need big datasets and high computing power regarding their operation. We definitely need to solve these problems to actually make the system work better and use it in real life.Chowdhery et al. [11], Zhai et al. [12], Li et al. [13].

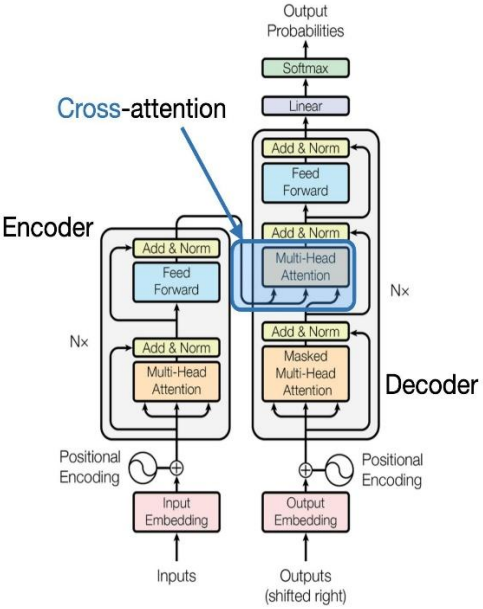

Figure 2: Cross-Modal Attention Mechanism in Multimodal Large Language Models

Table 2: Evolution of Multimodal AI Towards MLLMs

Stage	Approach	Description	Key Advancement	Limitation
Unimodal Systems	Text / Image / Speech Models	Processes single data type independently	Simplicity and specialization	No cross-modal understanding
Machine Learning	Feature-based Models	Uses handcrafted features for each modality	Data-driven learning	Feature dependency
Deep Learning	CNN, RNN, LSTM	Automatic feature extraction from raw data	Improved representation learning	Limited cross-modal interaction
Multimodal Learning	Fusion Models	Combines multiple modalities (text, image, speech)	Better contextual understanding	Data alignment issues
Transformer Models	Attention-based Models	Uses self-attention for sequence modeling	Captures long-range dependencies	High computation
MLLMs	Multimodal Transformers	Unified model for cross-modal reasoning	Integrated understanding across modalities	Resource intensive

Comparative Analysis

The shift from old methods to MLLMs surely shows major improvements in understanding different types of data. Moreover, this change highlights how much better these new systems are at handling multiple forms of information together. As per traditional methods, they are simple but do not understand the context regarding the situation. As per machine learning methods, adaptability gets better but regarding feature extraction, manual work is still needed. As per recent studies, deep learning models give better learning results but face problems regarding combining different types of data together. Further, as per recent studies, multimodal methods give better results by mixing different data types, while transformer models help understand context better regarding attention systems. MLLMs surely represent the most advanced stage in this field. Moreover, they offer unified frameworks that can perform cross-modal reasoning and provide integrated understanding across different types of data. However, these models surely create problems with computational complexity and scalability. Moreover, MLLMs represent the most advanced stage, offering unified frameworks that can handle cross-modal reasoning and integrated understanding. However, these models surely bring problems with computing complexity and scaling issues. Moreover, MLLMs represent the most advanced level, providing unified systems that can reason across different modes and understand them together. However, these models also bring problems with computer processing power and we are seeing issues with making them work for larger systems only. MLLMs

Research Gaps

Despite progress, further research gaps exist in multimodal understanding itself. We are seeing that current models only lack proper mixing of all three types - text, image, and speech - in one single system. Data alignment across different types actually faces big problems because structures and timing are definitely not the same.Basically, the fusion methods in most models cannot capture deep connections between different types of data - they all have the same limitation. Basically, getting large datasets with proper labels for multimodal systems is the same problem - there aren't enough available. These models actually need too much computing power, which definitely makes it hard to use them on a bigger scale.To fix these problems, we actually need to build better systems and definitely improve how different parts work together. We also need simple ways to train these systems properly. Liu et al. [14]

Summary

This literature review surely shows how multimodal understanding has changed from old methods to new transformer models and MLLMs. Moreover, it highlights the clear progress made in this field over time. We are seeing that each stage has only helped to make AI systems better at handling difficult data. As per current research, there are still problems regarding joining data together, making computers work faster, and handling large amounts of information. These areas need more study in future work.

Relevance to the Present Study

This study is further motivated by the limitations found in existing research itself. This work actually develops one framework that combines text, images, and speech using transformer models. It definitely aims to bring these different types of data together in a simple way. Basically, this study uses attention methods and smart data mixing to make the model understand context better and work the same way but with improved performance.Zhang et al. [15].

Table 3: Literature Survey of Multimodal Large Language Models for Integrated Text, Image, and Speech Understanding

S.No	Author & Year	Model / Approach	Modality	Key Contribution	Limitation
1	Lu et al. (2020)	VilBERT	Text + Image	Dual-stream transformer with co-attention	High computational cost
2	Tan & Bansal (2020)	LXMERT	Text + Image	Cross-modality encoder with multi-task learning	Complex training
3	Chen et al. (2020)	UNITER	Text + Image	Unified transformer for joint representation	Requires large data
4	Li et al. (2020)	Oscar	Text + Image	Uses object tags for better alignment	Limited to vision-language
5	Sun et al. (2020)	VideoBERT	Video + Text	Learns joint video-language representation	Limited scalability
6	Radford et al. (2021)	CLIP	Text + Image	Contrastive learning for image-text alignment	Data dependency

7	Ramesh et al. (2021)	DALL·E	Text + Image	Generates images from text	High computation
8	Jaegle et al. (2021)	Perceiver	Multimodal	Handles multiple modalities using latent attention	Complex architecture
9	Li et al. (2021)	ALBEF	Text + Image	Align-before-fuse strategy improves performance	Limited modalities
10	Jia et al. (2021)	ALIGN	Text + Image	Large-scale multimodal training	Requires massive dataset
11	Chowdhery et al. (2022)	PaLM	Text	Large-scale language reasoning	Not fully multimodal
12	Zhai et al. (2022)	CoCa	Text + Image	Combines contrastive + caption learning	High complexity
13	Li et al. (2022)	BLIP	Text + Image	Bootstrapped vision-language learning	Data quality sensitive
14	Yu et al. (2022)	BEiT-3	Multimodal	Unified transformer for vision-language tasks	Resource intensive
15	Tsimpoukelli et al. (2022)	Frozen LM	Text + Image	Uses pretrained LM with visual encoder	Limited adaptability
16	Driess et al. (2023)	PaLM-E	Multimodal	Integrates perception for robotics tasks	High computational cost
17	Huang et al. (2023)	Kosmos-1	Multimodal	Combines perception + language understanding	Limited speech integration
18	Liu et al. (2023)	VALOR	Text + Image + Speech	Tri-modal representation learning	Dataset dependency
19	OpenAI (2023)	GPT-4	Text + Image	Strong multimodal reasoning capability	Limited modality scope
20	Zhu et al. (2023)	Visual ChatGPT	Text + Image	Multimodal conversational system	Integration complexity
21	Peng et al. (2023)	InstructBLIP	Text + Image	Instruction-based multimodal learning	Training complexity
22	Liu et al. (2023)	MiniGPT-4	Text + Image	Aligns vision encoder with LLM	Limited robustness
23	Wu et al. (2023)	Survey Study	Multimodal	Comprehensive overview of MLLMs	No implementation
24	Wang et al. (2023)	Foundation Models	Multimodal	Large-scale multimodal architectures	High resource demand
25	Zhang et al. (2023)	Multimodal Transformer	Text + Image + Speech	Integrates tri-modal understanding	High complexity

IV. CONCLUSION

Multimodal Large Language Models surely mark a major step forward in artificial intelligence technology. Moreover, these systems can understand and work with text, images, and speech data together in one place. These models bridge the gap between human thinking and machine intelligence by providing context-aware reasoning capabilities that work across different modes. This further helps machines understand and process information like humans do, making the technology itself more intelligent. However, challenges of scalability and efficiency remain, and deployment itself in real-world situations needs further work. Future studies should surely work on making these models better and using them in different fields. Moreover, researchers must expand how these models can be applied in various areas of work.

REFERENCES

1. J. Lu, D. Batra, D. Parikh, and S. Lee, "VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 13–23, 2020.
2. H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 5100–5111.
3. Y.-C. Chen et al., "UNITER: Learning universal image-text representations," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 104–120.
4. X. Li et al., "Oscar: Object-semantics aligned pretraining for vision-language tasks," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 121–137.
5. C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2020, pp. 7464–7473.
6. A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 8748–8763.
7. A. Ramesh et al., "Zero-shot text-to-image generation," in Proc. Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 8821–8831.
8. A. Jaegle et al., "Perceiver: General perception with iterative attention," in Proc. Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 4651–4664.
9. J. Li et al., "Align before fuse: Vision and language representation learning with momentum distillation," Advances in Neural Information Processing Systems, vol. 34, pp. 9694–9705, 2021.
10. C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in Proc. Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 4904–4916.
11. A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," in Proc. Int. Conf. Machine Learning (ICML), vol. 162, 2022, pp. 5712–5730.
12. X. Zhai et al., "CoCa: Contrastive captioners are image-text foundation models," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 221–231.
13. J. Li et al., "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Machine Learning (ICML), 2022, pp. 12888–12900.
14. Y. Liu et al., "VALOR: Vision-audio-language omni-perception pretraining model," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 101–110.
15. P. Zhang et al., "Multimodal transformer for vision-language and speech understanding," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2023, pp. 1–6.