Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

Applications of Large Language Models in Multimodal Learning

YU, Peiyang [1*] XU, Xiaochuan [1] WANG, Jiani [2]

[1] Carnegie Mellon University, USA
[2] Stanford University, USA

**** YU, Peiyang is the corresponding author, E-mail: peiyangy@alumni.cmu.edu***

Abstract: In this paper, we provide a systematic review of the emerging field on applications for Large Language Models (LLMs) in multimodal learning, especially how such methodologies help improve orchestrated task performance by integrating different modalities like images, text, and audio. Multimodal learning is a field where we combine various types of data to make models learn multiple attributes and generate meaningful outputs. It is widely applied in image captioning, cross-modal retrieval, sentiment analysis, and speech recognition. It reviews the main multimodal learning approaches, such as feature extraction, modality alignment, and fusion strategies (early fusion, late fusion, and hybridization), and the performance of LLMs in cross-modal tasks. It highlights the present technological challenges, emphasizing concerns regarding computational resource utilization, model complexity, as well as a lack of multimodal fusion. Lastly, the article provides some suggestions for future applications on how to better integrate modalities and few-shot learning in cross-modal generation models. It also discusses ways to make multimodal machine translation systems run faster using less distributed computational power.

Keywords: Large Language Models (LLMs), Multimodal Learning, Cross-modal Tasks, Few-shot Learning, Cross-modal Generation.

Disciplines: Computer Sciences. **Subjects:** Large Language Models.

DOI: https://doi.org/10.5281/zenodo.14001455 **ARK:** https://n2t.net/ark:/40704/JCTAM.v1n4a13

1 1. INTRODUCTION

One of the key research areas in artificial intelligence today is multimodal learning. Its main objective is to combine the heterogeneous modalities of information (text, images and videos, sound etc.) for a holistic understanding and use. System can effectively combining these modes to enrich a large amount of contextual information will be able to perform task with higher accuracy and intelligence. The scope for multimodal learning has been expanded due to the large strides made by Large Language Models (LLMs) in efficiently utilizing text data, with models that are advancing rapidly and even dominating on this type of data. For instance, models like GPT-3 and BERT can not only understand and generate natural language but also drive intelligent applications such as image captioning, sentiment analysis, and human-computer dialogue through their combination with modalities like images and audio.

Technically, multimodal learning typically involves key steps such as feature extraction, feature alignment, and fusion strategies. The feature extraction phase utilizes deep learning models to extract meaningful features from each modality [1], while feature alignment ensures that data from different modalities can be effectively compared and operated upon in the same space. Fusion strategies determine how these features are integrated, with common methods including early fusion (fusion occurring after feature extraction) and late fusion (fusion taking place during the model decision phase). It achieves state-of-the-art performance levels in application cases such as medical image analysis, autonomous driving or social media content understanding. In the medical area, models could be more precise at predicting and diagnosing diseases [2] by leveraging routine data e.g. to prototype; in this case for procurement should combine from clinical text or imaging studies instead of only one. Last but not least, we are encouraged to address ethical and privacy concerns in the transparency of multimodal learning systems for more interpretable utilization as well as outperformance under low-data conditions with a wider scope of application focused on modality-related applications.

2 OVERVIEW OF MULTIMODAL LEARNING

The aim in multimodal learning is to complement machinelearning with a variety of data types. The basic tenets relate to what the modalities are, how they interact and

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

combine. Modalities usually correspond to various forms of data (words, images, sound and video) [3]. Interactions between different modalities in the problem are a very important thing for multimodal learning. A deeper understanding and effective utilization of these relationships can enable models to perform more comprehensively and accurately when handling complex tasks. Current research indicates that while multimodal learning has achieved certain results in areas like image-text pairing, sentiment analysis, and multimedia retrieval, it still faces numerous challenges.

Firstly, data alignment poses a significant challenge in multimodal learning [4]. The temporal and spatial alignment of different modality data often affects model performance. For instance, in video understanding tasks, accurately aligning video frames with corresponding audio or text descriptions is a key challenge. Temporal alignment involves synchronizing dynamic data, such as ensuring that audio changes correspond precisely with the content of video frames; spatial alignment requires different modality data to be effectively comparable and interactive within the same feature space. Addressing these alignment issues can not only enhance the accuracy of the model but also improve its ability to understand complex scenarios. Researchers might consider leveraging attention mechanisms from deep learning to dynamically weight data from different modalities, thus improving alignment effectiveness.

Secondly, modality fusion is a core aspect of multimodal learning [5]. How to effectively combine information from different modalities to fully utilize their advantages remains a bottleneck in current technological development. Common fusion methods include early fusion (combining features from different modalities during the feature extraction phase), late fusion (integrating results from each modality during the decision phase), and intermediate fusion (interacting at the feature level). Early fusion can effectively capture complementary information between modalities, but the increased feature dimensions may lead to training difficulties. Late fusion, while simpler, may overlook the interdependencies between modalities. Therefore, researchers need to explore more flexible and efficient fusion strategies, such as utilizing emerging technologies like Graph Neural Networks (GNNs) to facilitate dynamic interactions between modalities, thereby enhancing model expressiveness.

Third, model complexity is another issue that we cannot overlook in multimodal learning. With increasing number of modalities, the model becomes complex which makes training and optimization more difficult. Large models not only take more computational power to train but also tend to overfit as they contain so many parameters. This can be tackled by utilizing more efficient model architectures (e.g., lightweight neural networks) or quantization methods, which significantly reduce the computational costs and required resources for deploying this algorithm on real-world applications [6]. Also, transfer learning and self-supervised techniques also reduce the dependence on labelled data lowering our concerns around scarcity of dataset issues.

3 OVERVIEW OF LARGE LANGUAGE MODELS

Large Language Models (LLMs), such as BERT, GPT-3, and T5, represent a significant breakthrough in the field of natural language processing in recent years. These models are trained on vast amounts of textual data, enabling them to generate high-quality text, understand context, and perform complex reasoning. These capabilities make them essential in multimodal learning, with their relevance primarily reflected in the following aspects:

Firstly, text generation and understanding are key applications of LLMs in multimodal learning. LLMs can generate descriptions related to image content, thereby supporting image understanding. For instance, in image-text pairing tasks, LLMs can produce corresponding descriptions based on a given image, assisting computers in better understanding the image's content. This ability can be applied not only for image labeling but also in visual question-answering systems, allowing users to ask questions in natural language while the model leverages image information to provide relevant answers. Moreover, LLMs can serve as conditional generative models in image generation tasks, creating images that meet specified requirements based on textual prompts, further enhancing their potential in multimodal learning.

Secondly, modality transformation is another important contribution of LLMs in multimodal learning. LLMs facilitate the conversion between text and images, providing a foundation for multimodal applications. This capability allows different types of data to be effectively mapped and interacted with one another. For example, in cross-modal retrieval tasks, users can retrieve relevant images by inputting text descriptions and vice versa. This flexible modality transformation not only improves user interaction experiences but also opens up more possibilities for the generation and understanding of multimodal content.

Additionally, with technological advancements, researchers are exploring effective ways to integrate LLMs with other modality models, such as convolutional neural networks and image processing models, to further enhance multimodal learning performance. This integration can not only strengthen the model's ability to process different types of information but also promote information sharing between modalities, ultimately improving overall task execution outcomes.

4 APPLICATIONS OF LARGE LANGUAGE MODELS IN MULTIMODAL LEARNING

4.1 Combining Images and Text for Question Answering

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

In multimodal learning, the combination of images and text for question answering is a significant application area for Large Language Models (LLMs). The model needs to analyze the content of images, extract key information[7], and convert that information into natural language text answers. This process involves several technical steps, including image preprocessing, feature extraction, and language generation. Effective image analysis requires the model to recognize objects, scenes, and actions within the image and understand the relationships among them. At the same time, the model must integrate this visual information with its text reasoning abilities to provide accurate answers to complex questions. This capability relies not only on the model's visual understanding but also on robust contextual understanding and language generation abilities, ensuring excellent performance in practical applications.

4.2 Generating Descriptive Text Related to Image Content

Another important function of LLMs is to automatically generate descriptive text[8] based on the input image content. This capability helps users quickly grasp the themes and details of images, enhancing information retrieval efficiency. By combining image features with natural language generation techniques, the model generates natural language descriptions related to the images[9]. This generative ability is significant in content creation and social media publishing, contributing to improved user experiences and facilitating more natural and seamless human-computer interactions.

4.3 Sentiment Analysis and Recommendation Systems

Sentiment analysis and recommendation systems are crucial applications of LLMs in multimodal learning[10]. By combining user-generated text data (such as comments and ratings) with image data (such as product images), LLMs can analyze users' emotional tendencies, enhancing the accuracy of recommendation systems and overall user satisfaction[11]. This analysis aids in understanding user preferences and improving personalized recommendations, thereby providing more precise services. With advancements in big data and machine learning technologies, the integration of sentiment analysis and recommendation systems is expected to further enhance user experience[12].

4.4 Video Understanding and Speech Recognition

The application of LLMs extends to other multimodal scenarios, such as video understanding and speech recognition[13]. In video understanding, LLMs can analyze video content, generate corresponding descriptions, or extract key information to support video retrieval and classification. In speech recognition, by combining image or text data, LLMs can achieve more accurate speech transcription and intent understanding[14]. This integration not only enhances the interaction capabilities of voice assistants and other smart devices but also improves the overall user experience.

5 TECHNICAL ARCHITECTURE AND METHODS

In multimodal learning, effectively processing input features from different modalities is one of the core challenges for enhancing model performance[15]. The inherent heterogeneity of different types of data (such as text, images, speech, etc.) makes direct fusion challenging. Therefore, designing a reasonable technical architecture and fusion methods is crucial. Below are some commonly used multimodal fusion methods with in-depth discussions.

5.1 Early Fusion

Early fusion involves combining multimodal data at the input stage, specifically before the model begins formal training, after feature extraction. The idea behind this method is to transform data from different modalities into a unified representation in the same vector space, which is then directly input to the model[16]. This strategy allows the model to learn the collaborative relationships and complementary information between modalities from the early training stages, theoretically capturing deep interactions between them[17]. The main advantage is that the model receives complete information from multiple modalities simultaneously, offering opportunities to optimize the global representation of the multimodal input.

However, early fusion also presents several significant challenges. First, data from different modalities may have varying temporal scales, data distributions, and structures. For instance, image features are often high-dimensional pixel information, while text is usually represented as low-dimensional word vectors; direct fusion can lead to information asymmetry. Additionally, early fusion requires stringent data preprocessing, including standardization or dimensionality reduction for different modalities. Noise propagation is another risk; if the quality of one modality's data is poor, it can adversely affect the information transmission from other modalities. Nonetheless, early fusion is suitable for tasks where modalities have strong correlations and where capturing fine-grained interactions early on is crucial.

5.2 Late Fusion

Late fusion is a more separated multimodal processing strategy, allowing the model to independently process different modalities' data. In this approach, different modalities are processed through independent networks (or branches), generating their intermediate representations or predictions, which are then integrated at the output stage using methods such as weighted averaging, logistic regression, or other strategies. One major advantage of this

method is that the model can optimize each modality specifically, allowing independent processing without interference from other modalities.

Late fusion is particularly suitable for tasks with weaker correlations between modalities or when highly personalized processing is needed for each modality. For example, in certain multimodal medical applications, the feature distributions of images (such as X-rays) and texts (such as medical records) can differ significantly. Processing them separately before integration can enhance the overall system's robustness. Furthermore, since the representations of different modalities are generated independently, late fusion makes it easier to adjust each modality's contribution, such as through dynamic weighting or context-based fusion strategies.

However, the drawback of late fusion is that interactions between modalities are limited; the model cannot fully leverage the complementary information among modalities early on. This may lead to reduced information utilization efficiency in complex tasks, particularly those requiring deep integration of multiple modalities for fine-grained reasoning, such as visual-language understanding or multimodal dialogue generation.

5.3 Hybrid Fusion

Hybrid fusion combines the advantages of both early and late fusion, aiming to flexibly handle the information between different modalities. This method typically involves multiple stages of fusion throughout the model, integrating some modality information at the input stage while independently processing them in subsequent stages and then integrating again at the output stage. The main advantage of hybrid fusion is its flexibility, allowing for adaptive selection of appropriate fusion methods between different tasks and modalities, thus maximizing the complementarity of multimodal data.

A typical application scenario for hybrid fusion is in highly complex tasks, such as multimodal perception systems in autonomous driving. Autonomous driving needs to simultaneously process camera images, radar data, GPS information, and other sensor data. In such tasks, some modalities can be combined directly at an early stage (e.g., distance information from images and radar), while others (like GPS and map data) can be integrated at a later stage to provide a comprehensive understanding of the environment[18]. Hybrid fusion excels in such tasks as it can flexibly coordinate the interaction of modalities based on specific scenarios, ensuring that the model captures global information without overlooking local details.

5.4 Task-Specific Model Architecture Design

In addition to fusion strategies, designing optimized model architectures tailored to specific tasks is also critical in multimodal learning. For instance, the Transformer architecture, with its powerful global attention mechanism, has been widely applied to visual-language tasks in recent years. Through its self-attention mechanism, Transformers can capture complex associations between images and texts, performing exceptionally well in tasks like image description generation and cross-modal retrieval.

Another architecture of note is Graph Neural Networks (GNNs), which can represent relationships between different modalities through graph structures, particularly suitable for handling multimodal data with complex interactions[19]. GNNs learn the dependencies between modalities across the nodes and edges of the graph, capturing high-order semantic associations in the process.

In multimodal generation tasks, models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are also widely used[20]. They can map multimodal data into a shared latent space and generate new modal data through decoders. For example, VAEs can be used to generate images from text, while GANs can achieve the generation of corresponding textual descriptions from a given image.

6 EVALUATION METHODS

In multimodal learning, evaluation methods are key to measuring model performance, encompassing various metrics for tasks such as classification and generation. To accurately assess the performance of multimodal models, the following common evaluation metrics are typically employed:

6.1 Accuracy

Accuracy is one of the most fundamental and widely used evaluation methods, usually applied in classification tasks. It measures the proportion of correct predictions made by the model on a specific task. For example, in image classification or text classification tasks, accuracy reflects the overall predictive performance of the model by comparing its outputs with the true labels. Although accuracy is simple and straightforward, it may not provide a comprehensive evaluation on imbalanced datasets, and is sometimes used in conjunction with other metrics in practical applications.

6.2 BLEU Score

BLEU (Bilingual Evaluation Understudy Score) is a common metric used to evaluate specific modalities (typically text) in generation tasks. It measures the similarity between generated text and reference text, commonly used in machine translation and text generation tasks. By calculating the n-gram matching between the generated text and the target reference text, the BLEU score objectively reflects the quality of the generated text[21]. However, the limitation of the BLEU score lies in its heavy reliance on exact matches, which may fail to capture semantic similarities.

6.3 ROUGE Score

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is another widely used evaluation method for text generation tasks, especially for summarization tasks. Unlike BLEU, ROUGE focuses more on recall and measures the lexical matching between generated text and reference text. It evaluates the coverage of the generated content by calculating n-gram matches, word sequences, and the longest common subsequences. This method better captures whether the generated text effectively summarizes the main information of the reference text, making it widely applicable in automatic text summarization tasks.

6.4 Common Datasets

Common multimodal datasets include MS COCO (Microsoft Common Objects in Context) and Flickr30K, both of which play important roles in multimodal learning research.

MS COCO is a widely used dataset that encompasses a large number of labeled images and their corresponding textual descriptions. The dataset is designed to promote research in the fields of computer vision and natural language processing, particularly in image understanding and generation tasks. MS COCO provides over 300,000 images, each with multiple descriptions, allowing models to learn different textual expressions and semantic understanding. Due to its rich content, MS COCO is suitable not only for tasks like image classification and object detection but also provides strong support for image-to-text generation tasks. Models can enhance their cross-modal understanding by analyzing the relationships between image features and corresponding texts, thereby generating more accurate image descriptions.

Flickr30K is another important multimodal dataset focusing on image-text pairing tasks. This dataset contains over 30,000 images, each accompanied by five different descriptions, providing rich semantic information for model learning. Flickr30K is typically used for training and evaluating image-to-text generation, image retrieval, and other cross-modal tasks. Its design allows models to learn how to extract important features from images and transform these features into natural language descriptions, improving the model's performance in complex scenarios.

7 LIMITATIONS OF LARGE LANGUAGE MODELS

Despite the tremendous potential of Large Language Models (LLMs) in multimodal learning, enabling them to process and fuse data from various modalities with excellent performance, their application and development still face numerous limitations. These limitations are not only reflected in the high demand for computational resources and model complexity but also include deficiencies in modality fusion techniques. The following is an in-depth discussion of these issues.

7.1 Computational Resource Consumption

Large language models typically rely on substantial computational resources, particularly during the training and inference phases. This requirement arises because the models need to handle vast amounts of data and perform complex parameter optimization. For example, training a renowned LLM like GPT-4 necessitates hundreds or even thousands of GPU or TPU hours, leading to significant demands for hardware infrastructure and energy consumption. This high computational requirement directly limits the widespread application of LLMs, especially for small to medium-sized enterprises and research institutions that may find such costs prohibitive.

The issue of resource consumption extends beyond the training phase to the inference phase. In multimodal tasks, large models must process data from various modalities—text, images, audio, etc.—simultaneously, further exacerbating the demand for computational resources. For instance, in fields like autonomous driving and medical image analysis, hospital readmission[22], multimodal applications require real-time processing and decision-making. Traditional large models, with their longer computation delays and slower response times, struggle to meet the requirements of these time-sensitive tasks.

Therefore, future research should focus on model compression and optimization techniques to reduce reliance on computational resources. Methods such as Knowledge Distillation, Pruning, and Quantization can effectively reduce the model's parameter count and computational complexity, thereby decreasing resource consumption[23]. However, further exploration is needed to compress large models without sacrificing performance.

7.2 Model Complexity

As tasks become more complex and the number of modalities increases, the complexity of large language models also rises. This complexity manifests not only in the design of the model architecture but also impacts the training, debugging, and maintenance processes. Multimodal tasks often require differentiated handling of different types of inputs, placing higher demands on the model architecture. To accommodate the characteristics of various modalities, models may incorporate more branching networks, attention mechanisms, or interaction modules. While this multi-layered and multi-module design theoretically enhances model performance, it also makes effective management and debugging challenging.

The complexity of the model architecture complicates error diagnosis and troubleshooting. For example, when one modality of the model performs poorly in a task, it may stem from various factors, including data quality, feature extraction processes, inter-modality interaction mechanisms, or loss function design. The increased complexity of the

model makes problem identification more difficult, resulting in a time-consuming and labor-intensive correction process. Moreover, heightened model complexity may lead to overfitting, where the model excels on training data but performs poorly in testing or real-world scenarios. Thus, balancing performance with complexity presents a significant challenge when designing models for multimodal tasks.

Simplifying model architectures, introducing adaptive modular designs, and enhancing model transparency and interpretability are all important directions for future research. By designing more robust model structures and reducing redundancy and unnecessary complexity, it is possible to ensure performance while lowering the difficulty of management and debugging.

7.3 Deficiencies in Modality Fusion Techniques

Modality fusion is one of the core issues in multimodal learning; however, existing modality fusion techniques still exhibit significant shortcomings and struggle to effectively utilize information from all modalities. Different modalities of data may possess vast heterogeneity—for example, images are typically high-dimensional continuous data, while text is a discrete sequence of symbols. This disparity creates challenges in merging information from different modalities. Although methods such as early fusion, late fusion, and hybrid fusion are widely applied in multimodal learning, they often fall short in handling the complex relationships between modalities.

Current modality fusion methods frequently fail to adequately capture high-level dependencies between modalities. For instance, in cross-modal generation tasks (such as generating images from text or generating text from images), the associations between different modalities are often nonlinear and semantically complex, making it difficult for traditional fusion methods to grasp these subtle interactions. Furthermore, issues of information redundancy and noise are also pronounced during the modality fusion process. Certain modality information may not contribute meaningfully to the task but can interfere with information from other modalities, leading to a decline in overall performance. Additionally, in multimodal learning, the data from different modalities may be imbalanced—such as having more or more accurate text information than image information—creating challenges in addressing this modality imbalance during fusion.

To overcome these deficiencies, future research needs to develop more intelligent and adaptive modality fusion methods. For example, attention mechanism-based fusion strategies can dynamically adjust the weights of different modalities, flexibly determining which modality information should be prioritized. Moreover, Graph Neural Networks (GNNs) can be employed to represent the complex relationships between modalities, capturing high-order interactions through graph structures. These new fusion methods hold promise for enhancing the performance of multimodal models while minimizing information loss and interference between modalities.

8 FUTURE RESEARCH DIRECTIONS

Future research should focus on several key areas to advance the development of multimodal learning technologies and better support practical applications. First, the integration of information across different modalities remains a challenge, and research should develop more efficient fusion methods to enhance collaboration between modalities. Second, in practical applications such as healthcare, autonomous driving, and smart devices, accuracy and efficiency still need improvement, making it crucial to optimize algorithms and reduce dependency on computational resources. Furthermore, effective learning with limited data is an important direction, especially enhancing few-shot learning and cross-modal generation capabilities. As multimodal technology becomes more widely used, research should also address data privacy and system security issues to ensure that data can be processed safely and efficiently while protecting user privacy.

8.1 Improving Modal Fusion Technology

Current multimodal learning faces challenges in information fusion, particularly in how to effectively integrate information from different modalities. Future research should focus on developing more efficient fusion algorithms that retain the advantages of each modality while fully utilizing their collaborative potential. This may involve the application of attention mechanisms, self-supervised learning, and graph-based networks in deep learning to enhance information transmission and collaboration between modalities. Such improvements can not only boost performance in cross-modal tasks but also enhance the robustness and flexibility of models in handling complex scenarios. Expanding this research to include brain-computer interfaces (BCIs) opens new possibilities for multimodal learning. BCIs enable the direct integration of neural signals with external systems, introducing a novel and rich modality for data fusion[24].

8.2 Cross-Modal Generation Capabilities

Existing cross-modal generation models, such as generating images from text or vice versa are still limited in the fidelity of results and lack contextual details. In the future, it is necessary for further research to investigate how these models can be generated across more domains with better semantic cohesion (eg graphs of knowledge fine-grained) extended conceptual input and output generators multipart generative predominant admixtures internal adversarial networks advanced state-of-the-art parameterized decoders shared embedding variational autoncoders. Additionally, research should focus on making models generate content that is more natural and realistic, thereby improving

performance in cross-modal tasks and expanding application prospects in areas such as content generation, intelligent interaction, and virtual reality.

8.3 Few-Shot Learning

In the acquisition of multimodal data, particularly in scarce scenarios, few-shot learning has become a crucial research direction. Future research should focus on how to effectively utilize a minimal amount of data to achieve model generalization, including the development of frameworks based on transfer learning, meta-learning, or reinforcement learning to reduce reliance on large amounts of labeled data. Moreover, research should explore how to further mine and utilize the intrinsic correlations of multimodal data through self-supervised and unsupervised learning techniques to enhance model performance in resource-limited situations. This research direction will be significant for multimodal tasks that require large datasets, particularly in fields such as medical diagnosis, autonomous driving, and monitoring.

8.4 Optimizing Computational Resources

Being computationally and storage intensives, multimodal models are not well-suited to run on resource-constrained devices (smartphones, IoT). Therefore, an interesting research direction would be to use techniques such as model compression, pruning and quantization in order reduce the computational resources needed for a high-level of accuracy. In particular, it is easy for new compression algorithms and hardware acceleration technologies to greatly improve the efficiency of deep learning models in parameter optimization, memory usage and real-time processing. Meanwhile, research should also work on not sacrificing model performance for such optimizations and still enabling the models to perform well in resource-constricted environments. This is especially critical for promoting the application of multimodal technology in edge computing, IoT, and embedded systems.

8.5 Privacy Protection and Security

As multimodal learning is widely used in various intelligent applications, issues of user privacy and security have become increasingly important. Future research should delve into multimodal data privacy protection and model security, developing multimodal models with federated learning capabilities and differential privacy protection to ensure secure transmission and processing of data in distributed systems. Furthermore, it is essential to enhance research on defense mechanisms against adversarial attacks to prevent malicious manipulation of multimodal models. By constructing secure and reliable multimodal learning frameworks, the application potential in sensitive fields such as healthcare and finance can be significantly increased.

8.6 Multimodal Explainability and Explainable AI

As multimodal models become increasingly complex, their black-box nature has become more pronounced. Therefore, future research should focus on improving the explainability of multimodal models, making their decision-making processes more transparent. By developing more intuitive visualization tools and algorithms, users can better understand the decision logic of models across different modalities. This not only helps increase user trust in AI systems but also aids researchers and developers in debugging and optimizing models. Advances in explainable AI are especially crucial for fields that require high transparency, such as medical diagnosis and legal analysis.

9 CONCLUSION

The integration of large language models (LLMs) with multimodal learning demonstrates tremendous potential, driving advancements and applications in cross-modal tasks across multiple domains. With their exceptional capabilities in natural language processing, LLMs significantly enhance text understanding and generation in multimodal tasks, particularly in tasks like image captioning, sentiment analysis[25], and visual question answering, showcasing outstanding performance. However, despite considerable progress, the application of multimodal learning still faces numerous challenges. First, the consumption of computational resources remains a major bottleneck limiting the widespread application of large language models, especially during training and inference phases, where existing models often require substantial computational resources. This issue not only restricts usage by small and medium-sized enterprises but also introduces computational latency and energy consumption problems in applications with high real-time requirements. Second, as the number of modalities increases, the complexity of models escalates dramatically. The modality fusion and interaction strategies involved in multimodal tasks complicate model structures, increase the difficulty of training and debugging, and also lead to overfitting and performance bottlenecks. Third, current modality fusion technologies still have deficiencies, making it challenging to capture high-order dependencies and fine-grained information between modalities. Achieving more efficient interactions and information sharing between different modalities remains a challenge in current research. Future research should focus on the following areas: First, developing more efficient modal fusion strategies, particularly utilizing adaptive attention mechanisms and graph neural networks to enhance information transmission and collaboration between modalities. Second, enhancing few-shot learning and self-supervised learning capabilities to reduce reliance on large-scale labeled data, thus maintaining excellent model performance even in resource-scarce situations. Finally, research should further optimize the utilization of computational resources by employing techniques such as model compression, pruning, and quantization to reduce demands on hardware and energy. These improvements will lay the foundation for the

widespread application of multimodal learning, enabling it to play a larger role in fields such as healthcare, autonomous driving, and smart devices.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

YU, Peiyang

Information Networking Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, peiyangy@alumni.cmu.edu.

XU, Xiaochuan

Information Networking Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, xiaochux@alumni.cmu.edu.

WANG, Jiani

Department of computer science, Stanford university, Stanford CA 94305, jwang.tech@outlook.com.

REFERENCES

[1] Liu, W., Zhou, L., Zeng, D., Xiao, Y., Cheng, S., Zhang, C., ... & Chen, W. (2024). Beyond Single-Event Extraction: Towards Efficient Document-Level Multi-Event Argument Extraction. arXiv preprint arXiv:2405.01884.

[2] Zhang, Y., Wang, F., Huang, X., Li, X., Liu, S., & Zhang, H. (2024). Optimization and Application of Cloud-based Deep Learning Architecture for Multi-Source Data Prediction. arXiv preprint arXiv:2410.12642.

[3] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424-444.

[4] Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 56(10), 1-42.

[5] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424-444.

[6] Bačić, B., Feng, C., & Li, W. (2024). JY61 IMU SENSOR EXTERNAL VALIDITY: A FRAMEWORK FOR ADVANCED PEDOMETER ALGORITHM PERSONALISATION. ISBS Proceedings Archive, 42(1), 60.

[7] Liu, W., Cheng, S., Zeng, D., & Hong, Q. (2023, July). Enhancing Document-level Event Argument Extraction with Contextual Clues and Role Relevance. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 12908-12922).

[8] Leong, H. Y., Gao, Y. F., Shuai, J., Zhang, Y., & Pamuksuz, U. (2024). Efficient Fine-Tuning of Large Language Models for Automated Medical Documentation. arXiv preprint arXiv:2409.09324.

[9] Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., ... & Reitter, D. (2023). Measuring

attribution in natural language generation models. Computational Linguistics, 49(4), 777-840.

[10] Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports, 14(1), 9603.

[11] Scherrer, N., Shi, C., Feder, A., & Blei, D. (2024). Evaluating the moral beliefs encoded in llms. Advances in Neural Information Processing Systems, 36.

[12] Asaithambi, S. P. R., Venkatraman, R., & Venkatraman, S. (2023). A thematic travel recommendation system using an augmented big data analytical model. Technologies, 11(1), 28.

[13] Ataallah, K., Shen, X., Abdelrahman, E., Sleiman, E., Zhu, D., Ding, J., & Elhoseiny, M. (2024). Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. arXiv preprint arXiv:2404.03413.

[14] Marchisio, K., Ko, W. Y., Bérard, A., Dehaze, T., & Ruder, S. (2024). Understanding and mitigating language confusion in llms. arXiv preprint arXiv:2406.20052.

[15] Atrey, K., Singh, B. K., & Bodhey, N. K. (2024). Multimodal classification of breast cancer using feature level fusion of mammogram and ultrasound images in machine learning paradigm. Multimedia Tools and Applications, 83(7), 21347-21368.

[16] Quiles Pérez, M., Martínez Beltrán, E. T., López Bernal, S., Horna Prat, E., Montesano Del Campo, L., Fernández Maimó, L., & Huertas Celdran, A. (2024). Data fusion in neuromarketing: Multimodal analysis of biosignals, lifecycle stages, current advances, datasets, trends, and challenges. Information Fusion, 105, 102231.

[17] Atz, K., Cotos, L., Isert, C., Håkansson, M., Focht, D., Hilleke, M., ... & Schneider, G. (2024). Prospective de novo drug design with deep interactome learning. Nature Communications, 15(1),

[18] Wang, D. (Ed.). (2016). Information Science and Electronic Engineering: Proceedings of the 3rd International Conference of Electronic Engineering and Information Science (ICEEIS 2016), January 4-5, 2016, Harbin, China. CRC Press.

[19] Khemani, B., Patil, S., Kotecha, K., & Tanwar, S. (2024). A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. Journal of Big Data, 11(1), 18.

[20] Akkem, Y., Biswas, S. K., & Varanasi, A. (2024). A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Engineering Applications of Artificial Intelligence, 131, 107881.

[21] Ghassemiazghandi, M. (2024). An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score. Theory and Practice in Language Studies, 14(4), 985-994.

[22] Li, X., & Liu, S. (2024). Predicting 30-Day Hospital Readmission in Medicare Patients: Insights from an LSTM Deep Learning Model. medRxiv. doi:10.1101/2024.09.08.24313212

[23] Lu, J. (2024). Optimizing E-Commerce with Multi-Objective Recommendations Using Ensemble Learning.

[24] Liu T, Wu Y, Ye A, Cao L, Cao Y. Two-stage sparse multi-objective evolutionary algorithm for channel selection optimization in BCIs. Frontiers in Human Neuroscience. 2024 May 22;18:1400077.

[25] Yu, P., Cui, V. Y., & Guan, J. (2021, March). Text classification by using natural language processing. In Journal of Physics: Conference Series (Vol. 1802, No. 4, p. 042010). IOP Publishing.