International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Review of Recent Advances in Transfer Learning for Multimodal Data Analysis and Fusion

Anil Paila

B. E., Department of Mechanical Engineering,
Viswanadha Institute of Technology and Management
JNTU Kakinada University, Andhra Pradesh, India

ABSTRACT

The present analysis of the review focuses on the recent algorithms for transfer learning in data fusion and assessment. The first part is about transfer learning and its usefulness in connecting the data from other sectors to improve performance. The background section analyzes the many possibilities and limitations of multimodal data analysis and fusion. This review study discusses the domain adaptation, representation learning, multi-task learning and fine-tuned pre trained models as transfer processes. This is especially practical within the multimodal environment where text, illustrations, and audio recordings of voice and video content are gathered. In the review study, multimodal data analysis methodologies that are analysed include the multimode data, text-image fusion; audio and video understanding along with sensor datum fusion. In this part, the early, late and also hybrid fusion strategies are analysed with special consideration for attention mechanisms that significantly contribute to the fusion process.

Healthcare, finance, image/video processing, and also natural language processing are the classical applications of transfer learning in multimodal environments. In conclusion, the study identifies many problems and also suggests further research focusing on transfer learning as an integral part of the multimodal data fusion and analysis. This paper's approach, main findings and the field it covers are described in this abstract; readers should proceed to reading of an entire text that allows for a more profound understanding of such dynamic and rapidly changing phenomenon as cybercrime.

Keywords: Fusion, Multimodal Data, Hybrid Methods, Healthcare, Cybercrime

1. Introduction

Transfer learning is one of the machine-learning techniques that uses the knowledge obtained from a previous domain to improve the performance in another area. It has been defined as a very transformative approach. This review paper concentrates on "novel transfer learning approaches for the analysis, and fusion of multimodal data," attempts to analyse the development and integration of recent advancements in methods with respect to an advanced area so that merging the information originating from distinct sources is possible (Arnab et al., 2023).

The main two-fold objectives of this review are as follows. The aims of this work is to present a detailed understanding on transfer learning methods, and evaluate their effectiveness in complex multimodal setups where information inputs arise from variety of sources like text, images or sound video as well as sensed data (Rahate et al., 2023). In addition, the research focuses on new trends in methods for multi-modal data analysis and on combining different modalities clarifying the interrelations between such unique modalities.

This work seeks to address the escalating complexity and a range of contemporary data sources by using transfer learning in multimodal settings (Abdulnabi et al., 2018). It is important to be able, with the data world increasingly complex and rich in relevant information that such diverse forms of data can provide critical insights (Bandara and Patel 2021). The purpose of this research is to raise the reliability and effectiveness of the data analysis in complex, authentic environments through a comprehensive investigation into latest transfer learning algorithms, their utilization for multimodal analyses as well as fusion techniques (Baltrusaitis et al., 2019).

The relevance of this research lies in the fact that it promotes a higher level of proficiency in the multimodal data analysis and integration (Chen et al., 2021). The widespread influence of transfer learning is very noticeable in its wide implementation across many different fields such as healthcare, finance, image and also video processing among others. This research discusses the current trends and various issues with the application of machine learning methods to managing multimodal data. It functions as a great guide for many researchers and also practitioners, helping to shape the ongoing debate regarding how these different methodologies have changed over time (Cao et al., 2020).

The introduction serves as an excellent basis for the review paper and it also states its objectives, purposes, and a wider significance of researching transfer learning in multimodal data analysis and fusion (Ding et al., 2019). In this chapter, we will discuss different transfer learning techniques and their uses alot. This analysis will uncover the immense effects of these methods in dealing with the challenges presented by diverse and dependent data sources (Fernandez et al., 2018).

2. Transfer learning to the multimodal data analysis and fusion

To understand how the transfer learning has evolved within multimodal data analysis and fusion, one has to trace its history roots in machine learning and artificial intelligence (Gahrooei et al., 2019). Transfer learning emerged from the early machine-learning paradigms wherein models were normally trained and tested separately for isolated assignments. However, as the researchers faced the challenges of scalability and also adaptivity, there was an increasing need for methods that could leverage knowledge gained in one domain to improve performance across another (Bandara and Patel 2021).

Transfer learning first appeared in the computer vision industry together with CNNs. Pioneering efforts, such as the development of pre-trained models like ImageNet, showed that knowledge transfer from a huge picture classification task performed on one hand could help to improve the results in similar ones (Baltrusaitis et al., 2019). The first successes led to a desire for transfer learning outside in other fields, inspiring many scholars.

In the domain of multimodal data processing, its historical path is directly related to how different modalities have evolved a lot (Hong et al., 2021). One of the initial attempts in cross-modal research was to bridge the information between two modalities for better results on image captioning and sentiment analysis tasks (Chen et al., 2021). With technological breakthroughs in natural language processing and computer vision, the transfer learning approaches for audio video data study has become very optimal.

The historical background also reveals the growing importance of transfer learning as a much needed solution to address the many challenges arising from the vast amounts of and kinds of data (Baltrusaitis et al., 2019). As sensor technologies continue to become more ingrained in our daily lives, it has been further recognized by the academics that the potential of transfer learning lies in enhancing models' flexibility across many different kinds of data. This means that the use of more effective fusion treatments is indeed possible (Ding et al., 2019).

Transfer learning has evolved alot over the years, from simple one-modality applications to being highly revered in the multidisciplinary field of multi modal data analysis and fusion (Han et al., 2019). From the specialized models for a particular task to modern transfer learning among different modalities, this approach substantially prevails over the others. This prepares the ground for a detailed treatment of recent developments and challenges in the subsequent sections hereunder this review paper (Jiaxin et al., 2021).

Table.1. Description of the literature review of the existing work in last 5 years

S. No.	Area of study	Application	Challenges	Outcomes	Ref.
1	Multimodal Machine Learning	Systematic review of applications, challenges, gaps, and future directions	Data availability, model complexity, interpretability	Identified promising applications, highlighted challenges and gaps, proposed future directions	Arnab et al., (2023)
2	Pan sharpening	Fusion of high-resolution spectral and low-resolution spatial information	Spectral distortion, artifacts, computational cost	Proposed Hyper Transformer architecture for improved pan sharpening performance	Bandara and Patel, (2022)
3	Pan sharpening	Super-resolution of panchromatic images	Loss of high-frequency information, artifacts	Proposed Pan sharpening via Super-Resolution Iterative Residual Network with cross-scale learning strategy	Chen and Nan, (2021)
4	Remote Sensing Imagery Classification	Classification of remote sensing images using multimodal deep learning	Feature extraction, data fusion, model complexity	Demonstrated the effectiveness of multimodal deep learning for remote sensing image classification	Hong et al., (2021)
5	Remote Sensing Data Fusion	Review of deep learning methods for multimodal remote sensing data fusion	Data heterogeneity, model interpretability, computational complexity	Identified promising deep learning approaches for multimodal remote sensing data fusion, highlighted challenges and future directions	Jiaxin et al., (2022)

According to the analysis conducted by Arnab et al. in 2023, the swift and exponential growth of multimodal remote sensing data presents both advantages and difficulties for Earth observation operations. In recent years, tremendous progress has been made by combining their complementing characteristics. AI-related technologies have proven their supremacy in extracting features, giving them a greater advantage over traditional methods. The RS community is currently discussing the integration of multimodal RS data utilising DL-based techniques, particularly because to the availability of advanced methodologies and the volume of RS data described earlier. Hence, this review presents a comprehensive analysis of this swiftly growing discipline by scrutinising the relevant literature, methodically elucidating the integration of remote sensing across several prominent subfields, detailing the available resources, and forecasting its future expansion.

In their study, Abdulnabi et al. (2018) analyse the models, tasks, and data types employed in different fusion subdomains, with a particular focus on approaches based on deep learning. Overall, the utilisation of deep learning in combining many types of remote sensing data has yielded favourable results, suggesting the possibility of further extensive research.

Bandara and Patel (2022) have examined all aspects pertaining to multimodal deep learning. The research presents an intricate examination of various deep learning models across diverse media, subsequently categorising them based on their prospective applications. This article provides a comprehensive overview of numerous applications that utilise multiple modalities, surpassing earlier surveys on the same subject. These applications encompass a wide range of elements, such as images, music, video, text, body language, facial emotions, and physiological indicators, among others. We offer an innovative classification system for various multimodal deep learning applications, providing a more comprehensive and detailed analysis. The text also provides a concise overview of the evaluation measures, datasets, and architectures employed by these MMDL applications. Ultimately, we provide a comprehensive list of potential avenues for further research and highlight distinct areas of inquiry for each category of applications. The purpose of presenting this taxonomy and these research methodologies is to facilitate future researchers in the domain of multimodal deep learning to have a comprehensive understanding of the numerous unresolved inquiries within this topic.

The article by Chen et al. (2021) provides an overview of current trends in MML and examines its limits through a comprehensive literature review methodology. The study rigorously selected its subjects using the PRISMA methodology. 374 papers were chosen for further evaluation from a total of 1032 based on their successful fulfilment of the four specified research objectives. This review suggests that MML is influenced by modalities, ML algorithms, and workloads. According to the analysis, algorithms based on neural networks and data are the most widely used. This overview examines many elements of multimodal representation, translation, alignment, fusion, and co-learning, while also identifying potential deficiencies. Nevertheless, this SLR only presented a summary of MML's research. Further efforts are necessary to provide an abundance of future prospects for scholars intrigued by this interdisciplinary domain.

Fernandez et al. (2018) identified a significant problem with multimodal models: over fitting the training data, which reduces their generalizability. Small sample sizes are common in multimodal biomedical data sets due to the high cost and restricted availability of biological material used to construct them. When incorporating multi-omics data, the existing set of input variables typically expands, resulting in a larger number of variables. On the other hand, when representing various modalities, structures can possess a multitude of parameters. This may result in the training data exhibiting patterns that are not consistently beneficial.

Gahrooei et al. (2019) define “transfer learning” (TL) as the use of knowledge from one activity to enhance performance in a related task. One common method entails initialising the network's weights before training. By employing TL, the required sample size can be reduced by 50%. Further exploration of multimodal biological datasets is necessary to optimise transfer learning. Although fusion techniques incorporating TL exist, we believe that multimodal architectures utilising heterogeneous public unimodal datasets hold great potential for the future.

Hong et al. (2021) address the challenge of handling dynamic multimodal data. In the field of multimodal deep learning, it is well accepted that training a new model is necessary anytime there is a change in the data distribution. Nevertheless, the process of training such deep learning models is excessively time-consuming, and online applications that use several modes of input still fail to meet anticipated standards. Online and in-depth learning are two examples of synchronous approaches of acquiring new knowledge from fresh input while minimising the loss of existing knowledge. The emergence of dynamic multimodal data has emphasised the necessity of addressing the task of creating incremental multimodal deep learning models for online data fusion. The existing multimodal data is characterised by low quality due to the presence of noise, missing data, and outliers. Presently, numerous deep learning models are limited to processing noisy input from a single modality. Given the growing prevalence of low-quality multi-modal data, it is crucial to find a solution to the challenge of constructing a deep learning model that is suitable for such data.

2.1. Transfer Learning Techniques

Transfer learning techniques form the cornerstone of leveraging knowledge gained from one domain to enhance performance in another, providing a powerful paradigm in the context of multimodal data analysis and fusion (Summaira et al., 2021). This section explores key transfer learning methodologies that have demonstrated effectiveness across diverse modalities.

Domain Adaptation: Domain adaptation methods focus on aligning the feature distributions between the source and target domains (Bayoudh et al., 2022). By minimizing the distribution shift, these techniques enhance the model's ability to generalize from the source domain to the target domain. Domain adaptation is particularly relevant in multimodal scenarios where datasets may exhibit variations in terms of data types and modalities (Jiaxin et al., 2022).

Pre-trained Models and Fine-tuning: Pre-trained models, often initialized on large-scale datasets, serve as a valuable resource for transfer learning (Gahrooei et al., 2019). Models like BERT in natural language processing or pre-trained convolutional neural networks in computer vision capture generic features from vast datasets. Fine-tuning these models on specific multimodal tasks allows for efficient knowledge transfer, enabling the adaptation of learned representations to new data (Bayoudh et al., 2022).

Multi-task Learning: Multi-task learning involves training a model on multiple related tasks simultaneously (Han et al., 2019). In the context of multimodal data, this approach enables the joint learning of representations across different modalities. The shared knowledge learned from multiple tasks enhances the model's ability to handle the complexities of diverse data sources (Fernandez et al., 2018).

Representation Learning: Representation learning focuses on extracting meaningful and transferable representations from data (Gahrooei et al., 2019). In multimodal settings, learning common representations across modalities is crucial. Techniques like auto encoders and variational auto encoders are employed to capture latent features that encapsulate the shared information present in different data types (Li et al., 2018).

Understanding the nuances of these transfer learning techniques is paramount for effectively addressing the challenges posed by multimodal data (Fernandez et al., 2018). The selection of an appropriate transfer learning approach depends on the characteristics of the data, the task at hand, and the desired outcomes (Ramachandram and Taylor, 2017). The subsequent sections of this review will delve into the applications and advancements of these techniques in the realm of multimodal data analysis and fusion.

2.2. Multimodal Data Analysis

Multimodal data analysis involves the simultaneous exploration and interpretation of information from multiple, heterogeneous sources or modalities, such as text, images, audio, video, and sensor data (Soren et al., 2022). This section of the review paper focuses on recent advancements and key considerations in the realm of multimodal data analysis within the context of transfer learning (Song et al., 2019).

Text and Image Analysis: Combining textual information with visual data has been a prolific area of research (Wang et al., 2021). Transfer learning enables models to understand the contextual relationships between words and images, leading to improvements in tasks like image captioning, sentiment analysis, and cross-modal retrieval (Yi-Ming et al., 2023).

Audio and Video Understanding: Leveraging transfer learning in the domain of audio and video analysis enhances the comprehension of complex auditory and visual patterns (Ren, et al., 2022). Techniques such as transferring knowledge from pre-trained models on large audio datasets contribute to improved performance in tasks like speech recognition, sound classification, and video action recognition (Yokoya et al., 2017).

Sensor Data Fusion: In fields such as healthcare, environmental monitoring, and robotics, multimodal sensor data fusion is crucial (Soren et al., 2022). Transfer learning facilitates the integration of information from diverse sensors, enabling the creation of robust models that can handle the complexities of real-world scenarios (Kumar and Diwakar, 2021).

Multimodal data analysis is instrumental in capturing the richness and diversity of information available in modern datasets (Hong et al., 2021). Transfer learning acts as a catalyst, allowing models to extract shared representations from different modalities, leading to enhanced performance in tasks that involve the synthesis and understanding of information across multiple domains (Kumar and Diwakar, 2021). As the technological landscape continues to produce increasingly varied and interconnected data, the significance of multimodal data analysis, coupled with transfer learning techniques, becomes even more pronounced (Jiaxin et al., 2022). The subsequent sections of this review will explore fusion strategies and applications that leverage these advancements to address real-world challenges.

2.3. Fusion Strategies

Multimodal data analysis often involves combining information from diverse sources to extract comprehensive insights (Bayoudh et al., 2022). Fusion strategies play a pivotal role in integrating data from different modalities effectively (Summaira et al., 2021). This section explores key fusion strategies within the context of transfer learning for multimodal data analysis (Ramachandram and Taylor, 2017).

Early Fusion: Early fusion, or feature-level fusion, involves combining raw data from multiple modalities at the input level (Soren et al., 2022). The integrated features are then processed by a unified model. While this approach maintains the original information from each modality, challenges arise in handling disparate data types and scales.

Late Fusion: Late fusion, or decision-level fusion, involves training separate models for each modality, and their outputs are combined at a later stage (Song et al., 2019). This method accommodates diverse data types and allows individual models to specialize in their respective modalities, facilitating effective knowledge transfer (Yi-Ming et al., 2023).

Hybrid Fusion Techniques: Hybrid fusion techniques aim to combine the strengths of both early and late fusion (Ren, et al., 2022). This approach leverages shared representations at an intermediate level, allowing for the joint learning of features across modalities while maintaining the flexibility of modality-specific processing (Soren et al., 2022).

Attention Mechanisms: Attention mechanisms have gained prominence in multimodal fusion (Yokoya et al., 2017). They enable models to dynamically weigh the importance of information from different modalities, enhancing adaptability and robustness. Attention mechanisms play a crucial role in capturing relevant features while mitigating the impact of noise in the data (Fernandez et al., 2018).

Effective fusion strategies are vital for harnessing the complementary information inherent in multimodal datasets. Transfer learning acts as an enabler, facilitating the seamless integration of knowledge across modalities (Hong et al., 2021). Understanding the intricacies of fusion strategies is essential for designing models that can exploit the synergies between different data types, leading to enhanced performance in complex multimodal tasks. The subsequent sections of this review will delve into real-world applications that leverage these fusion strategies to address challenges across various domains (Summaira et al., 2021).

2.4. Design of heterogeneous networks

2.4.1. Fusion at the periphery of heterogeneity

Transforming heterogeneous data into vectors for better feature representation is a key advantage of multi-branch modality modelling (Abdulnabi et al., 2018). In order to facilitate more precise comparisons, these revised feature vectors can "level the playing field" across modalities with respect to data type, dimensionality imbalance, and scale. Similar to homogeneous intermediate fusion methods, classifiers can be fed such marginal representations (Baltrusaitis et al., 2019).

Bandara and Patel, (2022) integrated clinical data, laboratory testing, and information from CT scan margins. A combination of convolutional and recurrent neural networks forms the basis of the fusion model proposed by Jiaxin et al., (2022). It takes as inputs sequential clinical notes, time signals, and static admission and demographic data. Using the encoded static data, it augments latent feature spaces with the first two modalities. One possible outcome of this procedure is a patient representation suitable for use with a classifier.

As an example, feature selection can be used on the combined marginal representations to estimate the prognosis of patients with clear cell renal cell carcinoma. This involves identifying latent qualities that have the highest impact on the dependent variable. Kumar and Diwakar, (2021) decided against using additional hidden layers since the combined clinical and genetic data has low dimensionality. Still, the authors estimated that more clinical considerations might call for more joint hidden layers (Li et al., 2018).

As a general rule, when integrating data from several modalities, marginal heterogeneous fusion involves finding marginal representations for a subset of the modalities first (Gahrooei et al., 2019). When this occurs, the dimensionality curse does not affect the nonencoded modalities since they are low-dimensional. It follows that disentangled latent components may not be required to depict them (Bandara and Patel, 2022).

2.4.2. Fusion of diverse joint types

Because informative cross-modal interactions exist, it is plausible to believe that the various modalities do not affect the target variable independently. These relationships are expressed in joint heterogeneous intermediate fusion via feature interactions learned from marginal representations (Arnab et al., 2023). These interactions can be learned by combining marginal representations into a vector that is then fed into fully connected layers. Finally, an output layer that is task-specific is added (Bandara and Patel, 2022). For example, consider the following: Examples include using MRI in conjunction with clinical data to predict Alzheimer's disease and integrating MRI with genetic methods and clinical data to identify AD stages (Hou et al., 2019). Integrating latent representations from many imaging modalities with clinical data can also be used to predict the probability of liver transplantation for hepatocellular carcinoma (Chen et al., 2021).

Several scholars have contributed architectural modifications to the general concept for cooperative heterogeneous intermediate fusion in order to solve specific difficulties. Not collecting all modalities for each patient is a common issue in practice (Fernandez et al., 2018). Imputing missing modalities is difficult because training on entire samples reduces the size of the training set. Arnab et al., (2023) presented a multi-task network with task-specific output branches and unimodal input branches that may learn to deal with multimodal inputs lacking certain modalities (Bandara and Patel, 2022). Each action reflects one or more modalities. Training thus only affects the unimodal branch weights and task-specific branch weights. Additionally, homogeneous intermediate fusion can provide robustness to missing modalities, as demonstrated in (Fernandez et al., 2018).

Several approaches have been used to address the difficulty of making DL structures more interpretable in a multimodal setting. Gahrooei et al., (2019) achieved modality-specific interpretability by utilising genomic modalities such as convolutional, graph convolutional, and fully connected branches, as well as Grad-CAM for WSI and integrated gradient [98] for cell graph. Fernandez et al., (2018) used attention- and gradient-based interpretability in relation to WSI and molecular modalities. Furthermore, the various modalities are recognised for the contributions they make to forecast accuracy (Bandara and Patel, 2022).

Hong et al., (2021) validated gene expression prediction models using a multi-omics attention technique. These gradient- and attention-based strategies claim that models that allow for reliable biological interpretation remain invariant under heterogeneous intermediate fusion.

3. Applications

The integration of transfer learning and multimodal data analysis has paved the way for a myriad of applications across diverse domains, revolutionizing the landscape of information processing (Arnab et al., 2023). This section explores real-world applications that showcase the practical significance of leveraging shared knowledge from different modalities.

Healthcare: In healthcare, multimodal data analysis plays a crucial role in disease diagnosis, treatment planning, and patient monitoring (Rahate et al., 2022). Transfer learning aids in combining information from medical images, patient records, and sensor data, enhancing the accuracy of predictive models and personalized treatment strategies (Abdulnabi et al., 2018).

Finance: The financial sector benefits from transfer learning in tasks like fraud detection, risk assessment, and market analysis (Ding et al., 2019). By fusing information from textual news reports, financial data, and market trends, models can make more informed predictions and adapt to dynamic market conditions (Chen et al., 2021).

Image and Video Processing: Transfer learning enhances image and video processing applications by enabling models to understand complex visual patterns (Bandara and Patel, 2022). From object recognition in images to action recognition in videos, the ability to transfer knowledge across modalities improves the efficiency and accuracy of these tasks.

Natural Language Processing (NLP): NLP applications, such as sentiment analysis, language translation, and question answering, leverage transfer learning to understand the nuances of human language (Ding et al., 2019). Shared representations learned from diverse textual sources contribute to more nuanced and context-aware language understanding.

These applications underscore the versatility of transfer learning in multimodal data analysis, providing solutions to challenges in complex, data-rich environments (Gahrooei et al., 2019). As the world becomes increasingly interconnected, the ability to harness information from multiple sources becomes paramount across sectors (Han et al., 2019). The subsequent sections of this review will explore the evolving landscape of multimodal applications, shedding light on the innovative ways transfer learning continues to impact and transform various industries (Jiaxin et al., 2022).

4. Challenges and Future Directions

While the integration of transfer learning and multimodal data analysis has yielded significant advancements, several challenges persist, paving the way for future research directions.

Domain Shift and Heterogeneity: Addressing domain shift remains a prominent challenge, especially when dealing with diverse modalities (Ding et al., 2019). Ensuring that the knowledge transferred from one domain to another is applicable requires innovative techniques to handle heterogeneity in data distributions (Gahrooei et al., 2019).

Data Annotation and Availability: Multimodal datasets often demand substantial annotation efforts, especially when dealing with multiple modalities (Kumar and Diwakar, 2021). The scarcity of labeled data across all modalities can hinder the effectiveness of transfer learning techniques (Song et al., 2019). Future directions should explore strategies for efficient data annotation and the development of transfer learning methods with limited labeled data (Hou et al., 2019).

Intermodal Correlation: Understanding and modeling intermodal correlations present in multimodal data is crucial (Ramachandram and Taylor, 2017). Ensuring that the shared representations capture the complex relationships between different modalities requires advanced techniques, particularly in scenarios where modalities may be only loosely related (Gahrooei et al., 2019).

Adaptability to Dynamic Environments: Many real-world applications involve dynamic environments where data distributions can change over time (Wang et al., 2021). Developing transfer learning approaches that can adapt to such changes and remain effective in evolving scenarios is a critical area for future exploration (Bayoudh et al., 2022).

Ethical Considerations and Bias: As transfer learning models are trained on large and diverse datasets, ethical concerns related to bias and fairness emerge (Yokoya et al., 2017). Future research should address the ethical implications of transferring knowledge across domains, ensuring that models are unbiased and equitable across different demographic groups (Jiaxin et al., 2022).

Future directions in this field should focus on refining existing transfer learning techniques and developing novel methodologies to overcome these challenges (Soren et al., 2022). Additionally, exploring new modalities and combinations of modalities, such as incorporating temporal aspects in multimodal analysis, represents an exciting avenue for research (Chen et al., 2021). Robust evaluation metrics and benchmarks for multimodal tasks will also be crucial to objectively assess the performance of transfer learning models across different domains and applications (Fernandez et al., 2018). By tackling these challenges and embracing emerging trends, researchers can contribute to the on-going evolution of transfer learning for multimodal data analysis, making it even more robust and applicable to an ever-expanding array of real-world scenarios (Jiaxin et al., 2022).

5. Conclusions

As a result, the transfer learning methods applied to multimodal data analysis and fusion have ushered in an age where ML models are a lot more robust and flexible by combining information from multiple different sources. This review study presented a history of the transfer learning, multimodal data processing techniques for fusion as well as applications to diverse fields. As we look at the main findings of this exploration, some new themes and implications occur to us. Transfer learning has gone on from its first applications in images and text to become a very critical technique addressing the many challenges of contemporary data complexity and interdependence. Cross-disciplinary knowledge has allowed the health care, finance, digital image and also video processing as well as natural language understanding to progress.

Multimodal data can be analysed by the transfer learning – text, images, audio video and also sensor information. This holistic approach enables a deeper analysis and better understanding of the sophisticated real-life challenges. Fusion strategies such as early, late, and hybrid or attention are therefore very critical for the integration of data from various modalities. To select a fusion strategy based on the data qualities and also analytic objectives, an elusive approach is required. On the one hand, domain switchover, data annotation and intermodal correlation adaptively to transient situations and ethical issues belong among the challenges that have emerged in this discipline.

Transitional learning strategies should be developed in the further investigation, handling nonhomogeneous and dynamic data and ensuring the accuracy of predicted models. Creating relevant and appropriate evaluation standards and benchmarks will help this multidisciplinary discipline to develop. Herein, this review paper aims to summarize the transfer learning for multimodal data analysis and also fusion. By overcoming the challenges, embracing all trends and advancing in these field researchers may also make it very relevant in our data driven world.

References:

[1]. Arnab Barua, Mobyen Uddin Ahmed and Shahina Begum. (2023). *A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions.* IEEE Access, 10.1109/ACCESS.2023.3243854.

[2]. A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha. (2022). *Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions.* Inf. Fusion, vol. 81, pp. 203–239, May 2022.

[3]. Abdulnabi, A. H., Shuai, B., Zuo, Z., Chau, L. and Wang, G. (2018). *Multimodal recurrent neural networks with information transfer layers for indoor scene labeling.* IEEE Transactions on Multimedia, 20(7), 1656–1671.

[4]. Bandara, W.G.C. and Patel, V.M., (2022.) *Hyper Transformer: A textural and spectral feature fusion transformer for pansharpening.* arXiv preprint arXiv:2203.02503.

[5]. Baltrusaitis T. Ahuja C. and Morency LP. (2019). *Multimodal Machine Learning: A Survey and Taxonomy.* IEEE Trans Pattern Anal Mach Intell 2019; 41(2):423–43.

[6]. Chen, S., Qi, H. and Nan, K., (2021). *Pansharpening via super-resolution iterative residual network with a cross-scale learning strategy.* IEEE Trans. Geosci. Remote Sens. 60, 1–16. http://dx.doi.org/10.1109/TGRS.2021.3138096.

[7]. Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q. and Qiu, G. (2020). *Deep learning-based remote and social sensing data fusion for urban region function recognition*. ISPRS J. Photogramm. Remote Sens. 163, 82–97. http://dx.doi.org/ 10.1016/j.isprsjprs.2020.02.014.

[8]. Ding, R.; Li, X.; Nie, L.; Li, J.; Si, X.; Chu, D.; Liu, G.; and Zhan, D. (2019). *Empirical study and improvement on deep transfer learning for human activity recognition.* Sensors 2019, 19, 57.

[9]. Fernandez-Beltran, R., Haut, J.M., Paoletti, M.E., Plaza, J., Plaza, A. and Pla, F. (2018) *Multimodal probabilistic latent semantic analysis for sentinel-1 and sentinel-2 image fusion.* IEEE Geoscience and Remote Sensing Letters, 15(9), 1347–1351.

[10]. Gahrooei, M.R., Paynabar, K., Pacella, M. and Shi, J. (2019). *Process modeling and prediction with large number of high-dimensional variables using functional regression*. IEEE Transactions on Automation Science and Engineering, 17(2), 684–696.

[11]. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q. and Zhang, B., (2021). *More diverse means better: Multimodal deep learning meets remote-sensing imagery classification.* IEEE Trans. Geosci. Remote Sens. 59 (5), 4340–4354. http://dx.doi.org/10.1109/TGRS.2020.3016820.

[12]. Han, X.-H., Zheng, Y., and Chen, Y.-W., (2019). *Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution*. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 4330–4339. http://dx.doi.org/10.1109/ICCVW.2019.00533.

[13]. Hou, X., Sun, K. D., Shen, L., and Qiu, G. (2019). *Improving variational autoencoder with deep feature consistent and generative adversarial training*. Neurocomputing, 341, 183–194.

[14]. Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang and Jocelyn Chanussot. (2022). *Deep learning in multimodal remote sensing data fusion: A comprehensive review.* International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102926.

[15]. J. Summaira, et al. (2021). *Recent Advances and Trends in Multimodal Deep Learning: A Review.* College of Computer Science, Zhejiang University, China.

[16]. K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa. (2022). *A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets.* Vis. Comput., vol. 38, no. 8, pp. 2939–2970, Aug. 2022.

[17]. Kumar, P. and Diwakar, M. (2021). *A novel approach for multimodality medical image fusion over secure environment.* Transactions on Emerging Telecommunications Technologies, 32(2), e3985.

[18]. Li Y, Wu FX, Ngom A. (2018). *A review on machine learning principles for multi-view biological data integration.* Brief Bioinform, 2018; 19(2):325–40.

[19]. Ramachandram D, Taylor GW. (2017). *Deep multimodal learning: a survey on recent advances and trends.* IEEE Signal Process Mag, 2017; 34(6):96–108.

[20]. Soren Richard Stahlschmidt, Benjamin Ulfenborg and Jane Synnergren. (2022). *Multimodal deep learning for biomedical data fusion: a review.* Briefings in Bioinformatics, 2022, 23(2), 1–15, https://doi.org/10.1093/bib/bbab569.

[21]. Song, C., Liu, K. and Zhang, X. (2019). *A generic framework for multisensor degradation modeling based on supervised classification and failure surface.* IISE Transactions, 51(11), 1288–1302.

[22]. Wang, Y.; Feng, Z.; Song, L.; Liu, X.; and Liu, S. (2021). *Multiclassification of endoscopic colonoscopy images based on deep transfer learning.* Comput. Math. Methods Med. 2021, 2021, 2485934.

[23]. Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang and Zhi-Yuan Li. (2023). *Federated Learning on Multimodal Data: A Comprehensive Survey*. Machine Intelligence Research, www.mi-research.net, 20(4), 539-553, DOI: 10.1007/s11633-022-1398-0.

[24]. Y. Ren, N. Xu, M. Ling, and X. Geng. (2022). *Label distribution for multimodal machine learning*. Frontiers Comput. Sci., vol. 16, no. 1, pp. 1–11, Feb. 2022.

[25]. Yokoya, N., Ghamisi, P. and Xia, J. (2017). *Multimodal, multitemporal, and multisource global data fusion for local climate zones classification based on ensemble learning*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE Press, Piscataway, NJ, pp. 1197–1200.