International Journal of Machine Learning, AI & Data Science Evolution
ISSN: 3068-5073/ Open Access / IJMLAIDSE_Y1F3A005
Apex Academia Press / https://doi.org/10.63665/ijmlaidse-y1f3a005

Original Article

Combining Multi-Modal Data with Deep Generative Models

Rebecca John

Ladoke Akintola University of Technology

Abstract

In the past several years, AI tasks have had to deal with all sorts of data, such as text, audio, pictures, and sensor outputs. Multi-modal data fusion enhances one's insight into those factors by putting together the information coming from different data types. This work presents a unified framework for multi-modal data fusion using deep generative models, focusing on VAEs, GANs, and Diffusion Models. We present a novel architecture that can generate a joint latent representation capable of describing the relations among multiple modalities even if some data is noisy or missing. Extensive evaluations on benchmark datasets demonstrate that our approach outperforms state-of-the-art fusion methods on tasks such as classification, generation, and cross-modal retrieval. The proposed model extends advanced AI systems by showing their application to diverse fields, such as health care and multimedia content creation.

Keywords

Multi-modal learning, Data fusion, Deep generative models, Variational autoencoders (VAEs), Generative adversarial networks (GANs), Cross-modal generation, Representation learning, Diffusion models, Missing modality handling, Joint latent space.

Article History

Received: 16.07.2025

Accepted: 07.08.2025

Published: 14.08.2025

1. Introduction

A. Motivation for Fusion of Multi-Modal Data

In the real world, information rarely emanates from a single source. In determining how ill a patient is, identifying an object, or ascertaining how someone is feeling, one often relies on many types of data simultaneously. Sensor readings, sounds, pictures, and written descriptions coming together paint a full picture. Since that is complicated, we need systems that can put all this data together and make sense of it.

The way to combine these could be multi-modal data fusion. More than one source will help one learn more and learn better. For example, an autonomous car uses LiDAR sensors, GPS, and radar data with visual input to decide what to do. Multi-modal fusion will work towards enhancing tasks such as prediction, generation, and decision-making based on the best of each modality.

B. The Value of Combining Diverse Data Sources

Text tells you what things mean, audio tells you how things change over time, and sensor inputs often tell you what things are like right now. And for each kind of data, there's a different kind of information in that data. Systems that use more than one can make better, more valid decisions. Combining genetic data, a patient's medical history, and radiological images, for example, can make it much easier to find diseases and guess what will happen with them. Natural language processing works more like the way people understand things when you put text and pictures together, like in image captioning. Indeed, combining data from different sources is not only useful, it's often necessary in order to make AI systems that are strong and reliable.

C. Deep Generative Models' Function in Joint Distribution Learning

Deep generative models have been able to learn complex joint distributions, often nonlinear and with multi-modal data. Most of the old-style discriminative models can only perform predictions. However, generative models try to indicate the way data is distributed. Since they can grasp different modalities, capture interactions among them, and even generate one modality from the other, they are capable of dealing with some unique tasks, such as cross-modal translation, filling in missing modalities, and data synthesis. Some of the members of this

family of models that have been able to model such a distribution include GANs, VAEs, and more recently Diffusion Models. These models are good at solving problems relating to fusion and synthesis since they learn hidden representations showing the similarities between different modalities.

D. The Paper's Contributions and Scope

This paper aims to give a general framework of multi-modal data fusion using deep generative models and focuses in particular on VAE, GAN, and diffusion models. We first discuss the possibility of extending and adapting these models to various data modalities, even in the case of noisy or missing data.

We provide a unified theoretical framework explaining how conditional and joint distributions work across multiple modalities. Further, we propose an architecture that can be associated with hybrid, early, and late fusion strategies. We also establish training objectives such that the system guarantees handling incomplete data and functions effectively across different types of data. Finally, we evaluate our framework on several benchmark datasets and observe superior performance compared to state-of-the-art methods over recognition, generation, and retrieval tasks. This research will hopefully help us toward the development of newer, powerful, and flexible multi-modal AI systems.

2. Background and Related Work

A. Overview of Multi-Modal Data Fusion Techniques

There are three main categories of multi-modal data fusion: hybrid fusion, which combines the two; late fusion, which happens at a decision level; and early fusion, which happens at the feature level. Early fusion refers to the combination of low-level or raw features from different data types into a single model. It works well and is easy to use, but it assumes that all modes are always available and in sync, which is not often the case. On the contrary, late fusion operates on each modality separately and then combines the results at the decision level. Thus, it is easier to add and remove components without losing them. Hybrid fusion procedures aim to combine the benefits of both. Each strategy has positive and negative aspects related to effectiveness, complexity, and interpretability. Learning of joint representations is the new trend, given that deep learning models are much better at finding relations between different data types.

B. Deep Learning-Based vs. Conventional Fusion Techniques

In most cases, standard multi-modal fusion techniques, including decision trees, statistical correlation analysis, and kernel methods, were found to lack proper modelling of non-linear inter-modal interactions and to handle high-dimensional data. Deep learning gave way to more robust and adaptable methodologies for fusion. CNNs, RNNs, and transformers are some of the tools people have used to work with different kinds of data and combine their representations. Recently, attention mechanisms or autoencoders have been adopted to identify prevalent latent regions, enabling the dynamic assessment of each modality contribution. These newer methods are far superior and much more versatile than the earlier ones, particularly in recognizing speech, understanding how people feel in different modes, or even answering questions about pictures.

C. Introduction to Deep Generative Models (VAEs, GANs, Diffusion Models)

Deep generative models can generate new realistic samples and model the data distribution. VAEs learn a probabilistic latent space and then use the technique of variational inference to find the best lower bound on the likelihood. They are great at fixing things and filling in the gaps. GANs use adversarial training to generate data samples that look highly realistic. They are very good in fine sketching. These are the most recent types of models; these are called diffusion models. They take samples from a simple Gaussian distribution and progressively get rid of noise to learn more about how the data is distributed. All these have generated sounds and images that are really cool and new. You can use any of these architectures with multi-modal data by getting a joint latent representation or by making different types of data depend on each other.

D. Previous Research on Data Fusion Using Deep Generative Models

Deep generative models have been used for multi-modal tasks in a number of research. Cross-modal generation, such as creating images from text or vice versa, has been accomplished via conditional VAEs. In order to enable inference across modalities, even with partial inputs, joint VAEs learn a common latent space. GAN-

based models have been applied to problems such as audio-visual speech augmentation and text-to-image synthesis. In order to deal with missing modalities, several frameworks, such as MMVAE and MoPoE, train flexible priors. Even though their integration with multi-modal latent spaces is still in its infancy, diffusion models have recently started to be investigated for multi-modal fusion. These initiatives demonstrate the increasing interest in and promise of deep generative models for resolving the challenging multi-modal data fusion challenge.

3. Theoretical Foundations

A. Probabilistic Structure for Multi-Modal Combination

A probabilistic approach towards multi-modal fusion is employed to demonstrate the joint distribution, implying there is another way to do things for every xix_ixi in P (x1, x2, xn) P(x1,x2,.,xn)P(x1,x2,.,xn). This joint distribution or, rather, something very close to it is what we seek to obtain so that it may be used in applications involving inference, generation, or prediction in the case of unclear modalities. Introducing latent variables z which express the joint likelihood, we may write the generative process as P(x1,.,xn|z)P(z)P (x_1,. x_n ∣ z) P(z)P(x1,.,xn|z) P(z). This probabilistic approach offers a great avenue for filling gaps and making up deficiencies in data. This is crucial in understanding how much a kind of modality depends and has uncertainty about the others.

Table 1: Practical Role of Oracle Data Integrator (ODI) in Healthcare Analytics

Category	Explanation	Usefulness in Healthcare Analytics
Architecture Type	Uses ELT (Extract–Load–Transform) instead of traditional ETL. Transformations are executed inside the target system, reducing unnecessary data movement.	Improves processing speed for large clinical datasets (EHRs, imaging metadata, lab results).
Core Components	Includes a central repository for storing metadata/configurations and an execution engine for running integration tasks.	Ensures strong governance, auditability, and consistency—critical for healthcare compliance.
Supported Data Sources	Connects to relational databases, flat files, APIs, web services, cloud platforms, EHR systems, and LIMS tools.	Makes it easier to integrate heterogeneous hospital data without custom coding.
Performance Benefits	Uses target-system power for transformations, enabling high throughput and faster execution.	Allows near-real-time analytics for patient monitoring, risk scoring, and operational dashboards.
Scalability	ELT engine scales automatically with the target database or cloud environment's power.	Supports growing healthcare data volumes (IoT medical devices, wearables, remote monitoring).
Error Handling & Monitoring	Provides a graphical interface for debugging, flow monitoring, and dependency tracking.	Reduces risk of inaccurate clinical reporting and ensures data reliability.
Metadata Management	Centralized metadata handling for transformations, mappings, and business rules.	Ensures traceability of patient data lineage, supporting regulatory needs (HIPAA, GDPR).
Overall Benefit	Fast, flexible, visually manageable integration platform designed for complex data ecosystems.	Enables hospitals and research canters to build accurate, scalable analytics pipelines for clinical and operational decision-making.

B. Cooperative Representation Learning and Latent Variable Models

Examples of such latent variable models include VAEs and some variants of GANs. The unobserved variable, zzz, models how different types of data carry the same meaning. It will learn to merge the inputs from the different modes into this shared latent space such that it can put each mode back together. Joint representation learning hence allows for robust fusion to the extent that the latent space reflects shared information across different modalities. That is very useful in applications such as sorting or finding those things that need one representation to show all the data.

C. Across Modalities, Conditional and Joint Generation

Conditional generation is the process of turning one kind of data into another; for example, changing video into sound. In order for the process of generation to learn how to model P(xj|xi), a visible modality is required. On the other hand, joint generation is able to combine outputs from different modes in an intelligent way by taking samples from all modes at once from one latent variable z. In real-world applications, deep generative models do both conditional and joint generation. This helps when the modes may not be present altogether, or may be too loud.

4. Proposed Framework

A. Overview of Architecture

We propose a system that employs a multimodal generative architecture combining parts of VAE and GAN. It could also have parts that spread out to make a good generation. As can be seen, there are separate encoder and decoder networks for each type of data while there is also a shared latent space showing how all data is distributed together. We use modality-specific priors for more flexible generation, cross-modal attention layers to enable one modality to control how another modality learns its representation. The architecture is modular; that is, it should work with any combination of modalities and thus improve the cooperation between them.

B. Goals of Training

We optimize a combination of loss functions to ensure that the model learns meaningful and aligned representations. These include:

- Reconstruction loss, which guarantees that the latent space can be precisely used to rebuild each modality.
- GAN-inspired adversarial loss to promote realistic generation.
- Contrastive loss, which encourages modalities in the latent space to align discriminatively.
- Mutual information maximization, which makes sure that pertinent shared information across modalities is captured by the latent variables.

Our framework supports three fusion strategies:

- Early fusion: Raw inputs from all accessible modalities are concatenated at the feature level, then jointly encoded. Although synchronous availability and alignment are assumed, this is efficient.
- Late fusion: Provides robustness to missing modalities by processing each modality independently and then aggregating them at the decision or latent level.
- Hybrid fusion: A blend in which modality-specific branches are preserved and subsequently merged, while sharing some early-level characteristics. This strikes a balance between performance and flexibility. The choice of strategy depends on the task and the nature of the modalities involved.

C. Taking Care of Missing Modalities

Which Are Not There One of the nice things about our framework is that it can handle missing modalities both when it is training and testing. A previously proposed based on a product of experts (PoE) helps us do this. What this means is that even though one of the modalities is missing, the model still functions; it just omits one word from the product. We also employ stochastic modality dropout to make the model stronger by teaching it how to act as if it doesn't have information. For real-world use, it's important that the model can fill in missing modalities or make predictions based on any data it has.

5. Implementation and Experimental Setup

A. Utilised Datasets

We tried the proposed multi-modal data fusion architecture on a few benchmark datasets comprising various kinds of data. In this regard, we considered the MM-IMDb dataset, which contains pictures, sound clips, and written information about movies to help one find out what kind of movie it is and what people think of it. Compared to the original MNIST dataset, the AV-MNIST dataset was much improved since it comprised both audio and visual modes; it would show diverse ways of telling them apart. The CMU-MOSEI dataset contained a lot of movies containing words, pictures, and sounds that can give insight into what someone is trying to convey

and through what emotions. All these datasets provide us with all the information needed to know how well the proposed model works and its utilization for many more different tasks that use more than one mode.

B. Steps in Preprocessing for Various Modalities

These were pre-processed so that all types of data in these datasets worked together well and at their best. Pre-trained models like GloVe or BERT have been used to chop up the text and put it together in such a way that it makes sense to us. In visual modalities, convolutional neural networks were used which process data such that all the feature representations are of the same size. Programs like Open SMILE and COVAREP were used for feature extraction from audio and focused on prosodic parts like pitch and energy. Padding or interpolation was done in order to make all the modalities of the same length and in sync with each other. This careful preprocessing ensured that the multimodal fusion model got useful information from each different kind of data.

C. Details of the Model Architecture and Training

The design of model architecture takes pieces from Diffusion Models, GANs, and VAEs, puts them together, and makes them work better. It places the inputs into a common latent space using an encoder network operating internally on each type of input and then uses a decoder to reconstruct the original modalities from hidden representations. Cross-modal attention mechanisms are thus helpful for the model to learn more about how different types of data affect each other by having them collaborate. The model learns how to generate and align correctly across modalities using adversarial loss, reconstruction loss, and maximizing mutual information. We use Adam to find the best settings and a learning rate scheduler, and we stop early to avoid overfitting. This architecture and training schedule make it straightforward for the model to learn joint representations from input that comes from more than one source.

6. Results and Evaluation

A. Metrics That Are Quantitative

We went through plenty of numbers that show how well the suggested multimodal fusion model performs. Accuracy and F1-score have been used to find out the model's performance in sorting examples and finding an optimal trade-off between recall and precision. Metrics such as Inception Score and Fréchet Inception Distance were used with the purpose of checking the quality and diversity of the samples generated during the task performance. We also considered resilience and performance loss to find out how well the model could work without some parts. This set of numbers shows the model's performance in various situations for various tasks.

B. Qualitative Findings

Qualitative evaluation provides more than just numbers; it also depicts how the model performs in collecting data from various sources and arranging it together. We tried cross-modal generation-that is, pictures to written descriptions or text to audio-in order to see how well the results worked both visually and contextually. We prepared latent space visualizations using t-SNE or UMAP so that one can qualitatively see how the shared latent space is able to represent how different modalities work together. These qualitative evaluations indeed confirm the real-world applicability of the model through its capability for generating coherent and contextually relevant multimodal outputs.

C. Studies on Ablation

Several ablation studies were necessary in this line of thought regarding the contribution each component made to the proposed model. We systematically removed or changed parts-like cross-modal attention techniques, mutual information loss terms, or decoder architectures-to see how they would affect each performance metric. These showed ways to make things even better and how important every part is in getting optimum multimodality fusion. Among the findings was that one needs to plan in great detail, taking a lot of steps to discover how different types of transportation can work with one another.

D. Comparing with Cutting-Edge Techniques

Herein, we investigate the proposed model in combination with a number of sophisticated multimodal fusion techniques: early fusion, late fusion, and hybrid models. Comparing the results using performance metrics like F1-score, quality of generated text, and the accuracy of the model, the new model was found to outperform the

previous models, especially in handling missing modalities and generating better quality multimodal outputs. This is the best model in performing the assembly of cross-site information since it learned how to find links between information of different types and how to present them in useful form.

7. Applications

A. Medical Care

It is important that sensor data, clinical notes, and medical images be applied simultaneously in healthcare to ensure the patient's condition is completely monitored and diagnosed. The proposed multi-modal fusion model can combine diverse sources of data to come up with some high-value insights on personalized care, treatment planning, and disease prediction. Such a model will allow doctors to identify what is wrong with the patient and how his illness will progress by looking simultaneously at sensor data, medical history, and radiological images.

B. Self-governing Systems

It could be an autonomous car or a drone; in any case, the idea is that the system operates without intervention. They'll find their location and decide on an action depending on data from sensors and information concerning the outside world. Having proposed a model that can mix these modes using their own senses will help the system figure out what it has to do in complicated situations. Using LiDAR data and GPS coordinates combined with camera pictures, you are able to make maps accurately and find things on your way. In any case, the absence of people helping means that things will be much safer and more reliable.

C. Multimedia

Different types of multimedia applications require extracting data from diverse sources, such as content creation and recommendation systems. The proposed approach would be able to share knowledge across media types and transform the scripts into video and audio, or change the photos into text descriptions according to user preferences. Such an ability would make people happier and more engaged by providing them with information of value. For instance, suggesting songs or writing movie captions by what you see can make such a user experience even more fun.

8. Discussion

A. Benefits and Drawbacks

The advantages of the proposed multi-modal fusion paradigm include its ability to handle missing modalities, create high-quality multi-modal outputs, and make shared latent representations. On the other hand, shortcomings of the proposed model include requirements of a large amount of data and processing power for its functionality. It has complications that make understanding and quick decision-making difficult. Such aspects do require further consideration in relation to finding solutions for such limitations.

B. Real-Time Inference, Interpretability, And Scalability

That means when you make use of it for big projects, the idea has to grow. Ways of making the model smaller in size and using resources include knowledge distillation, quantization, and model pruning. One thing very important is knowing how this model makes its decisions; that is what it means for one to be able to comprehend something. This shall be helpful in making people believe in you and finding answers together. The two ways of learning more about how the model works are attention visualization and feature attribution. For example, self-driving cars have to make quick decisions and thus require real-time inference. Knowing how to make a model work better and faster will be very important in the future.

C. Biases in Data and Ethical Issues

Ethics become most important when using multi-modal fusion models, particularly in health and surveillance settings where privacy is very much essential. Honesty, fairness, and accountability are necessitated to stop hate and abuse. This should be very important because multimodal models, if they are trained on biased data, will make things even more unfair. For example, a health model trained to a great extent on the data of one group will not work well for those groups that aren't well represented. To make this less of an issue, one would want to utilize balanced methods and datasets for fair training. The application of tools which enhance privacy, like federated learning or differential privacy, will be in the interest of protecting user data. Model design and

application have to be transparent and honest to avoid findings of problems that shouldn't occur. We have to create and use such models so that they can be fair for everyone; this may enable everyone to improve.

9. Conclusion and Future Work

A. An Overview of the Contributions

This project brought together under one comprehensive framework elements of VAEs, GANs, and diffusion models for leveraging deep generative models in various types of data combination. We presented a new architecture that can learn, effectively, shared latent spaces from diverse types of data. Also, it is capable of effective generation and identification tasks. Our approach is suitable for real-world scenarios considering that it allows tackling missing modalities, enabling early, late, and hybrid fusion methods. We tested our method on many well-known datasets for multimodal data and found that it outperformed existing methods in quantitative and qualitative tests. We also spotted some other research questions about how to use our framework in healthcare, the creation of multimedia content, and self-driving cars, among others.

B. Unresolved Issues and Possible Research Paths

The results are good, but many things are yet to be fixed. First, apps with a lot of features and that happen in real time need to learn how to get better. Future work may investigate model efficiency enhancement strategies along with architectural design optimization. Second, it is a tough job, and it requires further research to learn how one can add more than three modalities at once without degradation in performance. Deep generative models need to be better interpretable for a number of reasons that are of importance. Another interesting direction is the combination of generative multi-modal models with external information sources or symbolic reasoning for enhancing cognitive capabilities. Further, there is a need to keep thinking about moral issues, especially those that concern privacy, fairness, and bias, as these models gain more favour.

10. References

[1] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ICLR*.

[2] Goodfellow, I. et al. (2014). Generative Adversarial Nets. *NIPS*.

[3] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *NeurIPS*.

[4] Ngiam, J. et al. (2011). Multimodal deep learning. *ICML*.

[5] Tsai, Y. H. H. et al. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. *ACL*.

[6] Wang, W. et al. (2020). Generalizing to Unseen Modalities for Multimodal Sentiment Analysis. *ACL*.

[7] Wu, Z. et al. (2018). Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*.

[8] Suzuki, M. et al. (2016). Joint multimodal learning with deep generative models. *ICML*.

[9] Shi, Y. et al. (2019). Variational Modality Dropout for Multi-Modal Deep Generative Models. *AAAI*.

[10] Saito, M. et al. (2017). Temporal Generative Adversarial Nets with Singular Value Clipping. *ICCV*.

[11] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*.

[12] Pu, Y. et al. (2016). Variational Autoencoder for Deep Learning of Images, Labels and Captions. *NIPS*.

[13] Radford, A. et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. *ICML* (CLIP).

[14] Ramesh, A. et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

[15] Tjandra, A. et al. (2020). Multi-modal self-supervised learning for audio-visual speech recognition. *ICASSP*.