2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET)

2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET) | 979-8-3315-4158-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/I3CEET61722.2024.10993792

An Overview Recent Trends and Challenges in Multi-Modal Image Retrieval Using Deep Learning

Aamir Khan
School of Computing
Graphic Era Hill University
Dehradun, India
aamircse36@gmail.com

Nisha Chandran S.
School of Computing
Graphic Era Hill University
Dehradun, India
nchandran@gehu.ac.in

D.R. Gangodkar
Dept. of Computer Science and Engineering
Graphic Era (Deemed to be) University
Dehradun, India
dgangodkar@yahoo.com

***Abstract*— Recent years we have seen a major increase in interest in multi-modal image retrieval, which includes looking for and getting pictures based on many modalities including visual, textual, and audio information. This review paper gives a general overview of current trends and challenges in multi-modal image retrieval, with an emphasis on the developments made possible by deep learning methods. We examine several methods, designs, and datasets used in multi-modal image retrieval, highlighting the potential advantages and emerging challenges in this area. This paper offers a thorough examination of present innovations and challenges in multi-modal image retrieval, with an emphasis on the developments made possible by deep learning methods, evaluation metrics, architectures, and issues related to multi-modal image retrieval. We also provide a comparative analysis in tabular form to show the benefits and drawbacks of various approaches.**

Keywords*— Multi-Modal Image Retrieval, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, Attention Mechanisms, Evaluation Metrics.*

I. INTRODUCTION

In the digital age, the ever-increasing volume of multimedia content has led to a surge in demand for efficient methods of image retrieval. Image retrieval is a fundamental task in computer vision, allowing users to search and access relevant images from vast databases or the internet. Traditional image retrieval approaches primarily rely on textual annotations or visual features extracted from single modalities, limiting their capability to capture the rich and diverse information present in multi-modal data. To address these limitations, the emergence of deep learning techniques has revolutionized the field of image retrieval, enabling the development of more powerful and flexible retrieval systems.Multi-modal image retrieval involves retrieving images by leveraging information from multiple modalities, such as text, visual content, and audio or even sensor data. Multi-model image retrieval, on the other hand, focuses on retrieving images across different domains or modalities, for example, retrieving images from a dataset of paintings given a query photo. These tasks present unique challenges due to the heterogeneity of data sources and the complexity of learning joint representations across modalities.

Recent years have witnessed substantial advancements in multi- modal image retrieval using deep learning techniques. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models have been at the forefront of these developments. The ability of deep learning models to automatically learn high-level feature representations has significantly improved the accuracy and scalability of image retrieval systems, making them more suitable for real-world applications. In this paper, we aim to explore the recent trends, methodologies, and challenges in the field of multi-modal image retrieval using deep learning. We will also focus on architectures of state-of-the- art models and investigate the various approaches employed to use and learn representations from diverse data modalities. Furthermore, we will discuss the applications of such systems in real-world scenarios, including image retrieval system, content- based recommendation systems, and image-to-image retrieval tasks.[1].One of the most important model named deep neural networks(DNNs)that's have perform multiple tasks that wide function are focus on image captioning ,segmentation and object identification. Traditional image retrieval methods based on textual or visual features often fail to capture the complexity and semantic information contained in images [2]. As a result, the integration of multiple modalities through deep learning models has become increasingly popular for accurate and comprehensive image retrieval. In next we chapter we discussed about the multi-modal image retrieval method or techniques that were used in the same domain.

II. MULTI-MODAL IMAGE RETRIEVAL TECHNIQUES

Multi-Modal Image Retrieval Techniques are sophisticated approaches and algorithms for retrieving relevant pictures from a database by integrating and analyzing several data modalities such as text, audio, and visual aspects[8]. The major objective is to bridge the information barrier in order to get more accurate and complete image retrieval. The term "multi-modal" refers to the combining of many sources of information, each of which captures a distinct part (attributes) of an image's content. These machine learning modal allow users to do picture searches utilizing a variety of input, such as written descriptions, voice inquiries, or even other images [9]. Multi-modal image retrieval approaches can improve the efficiency and precision of image search systems by integrating and exploiting information from diverse modalities. To perform multi-modal image retrieval, advanced technologies such as multi-modal retrieval, fusion- based algorithms, deep learning models, and attention processes are used. Multi-modal retrieval allows searches and results to move between data kinds, whereas fusion-based approaches integrate data from several sources into a single representation

[5]. Deep learning models use neural networks to extract features and discover similarities, allowing for effective multi-modal analysis. These methods have a wide range of applications, including multimedia retrieval, content-based picture retrieval, and cross- modal search. Multi-modal image retrieval techniques serve a critical role in enhancing image retrieval technology and enriching user experiences across varied areas by enabling more natural and efficient interactions between users and picture databases. Now we explains some modality where an image can be retrieval with any of them feature modality. So at first we discussed visual modality.

A. ***Visual Modality :*** The implementation of visual signals and information as a means of communication, perception, and comprehension is referred to as Visual modality [6]. It is one of the key ways humans and many other living organisms connect with their surroundings to acquire information. The visual modality in the context of multi-modal image retrieval techniques entails acquiring and processing visual material from pictures. Color, shape, texture, and spatial layouts of items within a picture are all examples of visual information [7]. The extraction of these characteristics allows for the production of visual representations, also known as feature vectors, which serve as numerical representations of the visual information of an image. The visual modality is critical in computer vision and image processing jobs. It allows machines and algorithms to comprehend and interpret visual data, allowing them to perform tasks such as picture classification, object identification, image segmentation, and content-based image retrieval. The visual modality is frequently used in a variety of applications, such as driverless cars, surveillance systems, medical imaging, and multimedia content analysis [1].

(i)Convolutional Neural Networks (CNNs): CNNs are used in multi-modal image retrieval to retrieve visual features. Convolutional Neural Networks (CNNs) is an example of deep learning model that has transformed computer vision and image recognition [6]. All Algorithms are design by real organisms' visual processing and have shown to be extremely successful in tasks like as picture categorization, object identification, segmentation, and more. Because of its capacity to automatically learn and discover patterns within visual data, CNNs are particularly well-suited for applications requiring grid-like data, such as photographs [3].

(ii)Attention Mechanisms in Visual Modality*:* The visual modality uses attention mechanisms that are modeled after human visual attention mechanisms [13]. In order to enhance the performance of tasks involving visuals, they are frequently utilized in deep learning models, particularly in the context of convolutional neural networks (CNNs) and transformer-based architectures [7]. Similar to how people choose attend to certain items or regions in a scene, attention methods allow the model to focus on key parts of the input data while overlooking less important regions [13].Attention mechanism is basically working with some predefine attributes such as Self-Attention in Transformers, Spatial Attention in CNNs, Channel Attention in CNNs, Guided Attention and Saliency Maps etc. The field of artificial intelligence (AI) has advanced significantly with the integration of attention processes into the visual modality. Significant advancements in a number of essential duties, including image classification, object identification, image segmentation, and image synthesis, have been made as a result of these attention techniques. Attention mechanisms have unlocked increased efficiency and precision in the management of complex visual patterns by allowing models to carefully allocate processing resources and concentrate on the most pertinent information. Models that generalize more well across various datasets and have greater interpretability as a result of this increased focus on important features are more transparent and easy for humans to grasp. As attention mechanisms develop and become more widely used in deep learning systems, their contribution to improving the performance and interpretability of visual tasks becomes more important.

B. ***Textual Modality:*** The term "Textual Modality" in the context of digital image Processing and the scientific community refer to the unique modality or type of information that is contained in an image that is in the form of text. If we take it into different way, the textual modality refers to the processing and analysis of images that contain text, with the main goal being to extract, identify, and comprehend the textual material that is contained in the image. A significant area of study in the larger field of computer vision and image processing is textual modality. According to our literature study we found that it is used last 20 years for various field and application such as Optical Character Recognition (OCR), Document Analysis, Scene Text Recognition, Handwritten Text Recognition, Text Detection and Localization, Language Translation and Understanding. Researchers in textual modality are dedicated to creating strong algorithms capable of handling diverse fonts, styles, and languages using machine learning, deep learning, and computer vision. These advancements have been instrumental in bridging the gap between visual and textual realms, facilitating intelligent analysis of text-containing image data.

1) ***Recurrent Neural Networks (RNNs) for Textual Modality:*** A number of reasons, including its capacity for handling sequential input and capturing temporal relationships, recurrent neural networks (RNNs) have been extensively employed in the textual modality of natural language processing. RNNs excel at jobs involving text when comprehending the meaning of a phrase or document depends on the sequence and context of the words. Based upon Some important characteristics of RNNs in textual modality, We can say that this technology have various role like Sequential Processing , Memory Retention , Vanishing and Exploding Gradients , including language modeling application, sentiment analysis , machine translation field, text generation, and named entity recognition in image.

2) ***Joint Embedding of Visual and Textual Modalities****:* A multimodal learning approach called joint embedding of visual and textual modalities combines data from both visual (images or videos) and textual (natural language) sources into a common representation space. The objective is to document the semantic connections and linkages between visual and textual data in order to provide a more complete and integrated understanding of the multimodal data. By enabling multi-modal retrieval, multimodal fusion, and effective interactions between visual and textual information, this integrated embedding space enhances performance in a variety of tasks, including picture captioning, visual question responding and multi-modal knowledge retrieval. Several methods, including multi-modal matching loss functions,

Siamese networks, and attention mechanisms that capture the links between modalities during training, are used to achieve the combined embedding of visual and textual modalities. The Visual Semantic Embedding (VSE), the Multimodal Compact Bilinear Pooling (MCB), and the Vision-and-Language Pre-training (e.g., CLIP) are well-known models that use joint embedding for multimodal learning.

(3)Word Embedding's and Textual Representations: Textual presentations and word encoding are essential elements of Linguistics (NLP) that transform textual input into numerical forms appropriate for machine learning application and models. While textual representations capture the semantic context of complete texts, word embedding to other textual data sets after that model will effectively convert into word-level representation. Natural language processing (NLP) relies heavily on the notions of word embedding and textual representations to effectively interpret and process human language. They are methods for transforming textual data into useful numerical vectors for machine learning models. In this process we mainly used two conversation functions that is Word Embedding, Textual Representations they are model based in entire mechanism.

C. Audio Modality: The term "audio modality" describes how sound data is represented and processed, including speech, music, and ambient noises. Audio modality is the process of turning audio impulses into numerical representations in the context of digital signal processing and machine learning. These enable machines to analyses and comprehend the content and context of audio data [14].Traditional audio processing can take place in either the digital or analogue domain, but because digital signal processing techniques are so much more potent and effective, digital representations are preferred by the majority of current audio systems[15].Audio modality will be utilized with combination with other, such as visual and textual data.

1) **Audio Representations and Features:** Audio representations and features are essential components in audio signal processing and machine learning tasks involving audio data. They are numerical representations that capture relevant information from audio signals, enabling efficient analysis and understanding of the audio content. In this approach we have various common models that helpful for audio modality representation such as Time-domain waveform, Spectrogram, Mel-Frequency Cepstral Coefficients, Chroma Features, Mel-Spectrogram, Log-Mel Energies, Zero Crossing Rate, Pitch and Timbre etc.

2) ***Audio Processing Techniques:*** Audio processing methods are essential for managing audio data properly and making it possible to retrieve information across several modalities, including pictures, texts, and videos. In order to represent audio material with other modalities in a common embedding space, audio processing techniques are utilized to extract useful properties from audio signals. To represent audio data in a numerical manner appropriate for multi-model-retrieval tasks, audio processing techniques are used. The frequency and time-domain properties of the audio signals are captured using common audio representations including spectrograms, log-mel energies, and Mel-frequency cepstral coefficients (MFCCs). These representations aid in distilling the audio data into concise and detailed feature vectors. Multi-modal fusion approaches can be used when audio data is modeled in the common embedding space. Through the use of these approaches, a thorough multimodal representation is produced by combining representations from several modalities.

3) ***Fusion of Visual, Textual, and Audio Modalities:*** Multimodal fusion usually referred to as the fusion of visual, textual, and auditory modalities, is an important idea in the field of multimodal learning. To provide a thorough and potent representation of the underlying data, it entails combining information from several data sources, including photos, texts, and audio signals. Utilizing of Multimodal fusion seeks to take use of each modality's capabilities, boost performance overall, and better comprehend and analysis complicated, real-world data. To know about how the fusion mechanism is working to visual, audio and text we have various pre define methods such as Representation Extraction & Modality-Specific Representations, Multimodal Learning Tasks etc. These are the methods that have capability to migrate and combine the different data in a multimodal platform.

III. DEEP LEARNING ARCHITECTURES FOR MULTI-MODAL IMAGE RETRIEVAL

According to various literature studies we found that deep learning architectures (DLA) are basically multi-model neural network that specifically works for or designed to retrieve images from the well-structured data base. We have various multiple modalities that are based on query. DLA is combine knowledge from various sources in the form of text, audio, visual to create a shared embedding space, after that corresponding modalities are represented as vector data with their semantic meaning that closer to each other [16].

A. Siamese Networks and Triplet Networks: A class of neural networks known as "Siamese Networks" is created for similarity-Based tasks like image similarity or "one-shot learning"[18]. One meta-learning model that has recently proven effective in using FSL (few-shot learning) in a variety of domains is the Siamese network. A specific network design used to generate and assess feature vectors for each input is called a siamese network. They consist of two or more comparable sub networks that have been setup with the same settings and weights [19] It is feasible to ascertain how similar the inputs are by contrasting the feature vectors of the Siamese network and training a similarity function between labeled points. This makes use of the Triplet Siamese Network, which has three identical sub networks, to classify COVID-19 patients with minimal training CT scans [19]. We used a group of fine-tuned pre-trained CNNs and adjusted the network weights using the pairwise margin ranking loss function. To produce an accurate and trustworthy model for generating predictions, assembling integrates many models [8].

B. Deep Cross-Modal Retrieval: Deep Cross-Modal Retrieval is an essential and challenging research area that aims to enable effective information retrieval across different modalities using deep learning techniques. It has numerous real-world applications and continues to be an active and evolving field of research. It describes the process of utilizing deep learning algorithms to search and retrieve data from many modalities (such as photos, text, and audio). The focus of conventional information retrieval is often on finding information using the same medium. For instance, the system

may extract pertinent text documents from a text query. However, retrieving data from several modalities is necessary in many real-world applications. For instance, the system should return pertinent text descriptions in response to an image query, or vice versa. By utilizing the strength of deep learning models on image processing, which have demonstrated exceptional effectiveness in a variety of AI tasks, Deep Cross- Modal Retrieval seeks to overcome this problem. Deep neural networks are being used by academics in an effort to learn a common representation space where data from many modalities may be closely mapped. According to our perception the primary goal of Deep Cross-Modal Retrieval in research is to develop algorithms that can effectively bridge the semantic gap between different modalities and perform efficient and accurate retrieval.

C. Generative Adversarial Networks (GANs) for Multi-Modal Retrieval: In research, Generative Adversarial Networks (GANs) have been thoroughly investigated and used for Multi- Modal Retrieval. GANs are a subclass of deep learning models made up of the generator and discriminator neural networks, which are trained concurrently and in competition. While the discriminator tries to tell the difference between genuine data and created data, the generator strives to create realistic data samples. This adversarial training method teaches GANs how to produce high-quality samples that closely match real data. GANs have been used in the framework of Multi-Modal Retrieval to close the gap between different modalities and develop a common representation space where data from many modalities may be successfully matched and compared. Further GANs basically work with field where complex problem can be fixed, that is Cross-Modal Generation, Joint Embedding Learning, Image-to- Text and Text-to-Image Retrieval, Cross-Modal Translation etc. It's crucial to remember that training GANs may be difficult, and that achieving steady and successful convergence continues to be an important study field. Additionally, getting large-scale multi- modal datasets for training GANs in these situations might be difficult.

IV. DATASETS FOR MULTI-MODAL IMAGE RETRIEVAL

The datasets used are primarily associated with its susceptibility to highly imbalanced issues. The challenge of imbalanced data arises when training a deep learning model for complex tasks [5]. This imbalanced dataset often leads to a biased or skewed prediction, impacting the model's performance [17] [19] [20].

A. MSCOCO (Microsoft Common Objects in Context):

MS-COCO is a widely used dataset for image-related tasks. It contains images with multiple annotations, including object detection, image segmentation, and image captioning [20]. MS-COCO is a widely used Dataset for various computer vision tasks, including image captioning and image retrieval. It contains more than 123,000 images, each paired with at least five descriptive captions [20].

B. **Flickr30K:** A well-liked benchmark for photo representation using sentences is the Flickr30k dataset. Over 31,000 photos make up the dataset [21]. There are five reference sentences given by human annotators for each image in the collection [22]. Each picture's annotations enable advancements in grounded language interpretation and automated image description [21] [22].Each image is associated with five descriptive sentences provided by human annotators.

C. **CUB-200-2011 (Caltech-UCSD Birds-200-2011):** The Dataset known as Caltech-UCSD Birds-200-2011 (CUB-200-2011) is the one that is most frequently used for tasks requiring fine-grained visual classification [23]. 11,788 photos, 5,994 for training and 5,794 for testing, of 200 bird subcategories are included with each image annotated with bounding boxes, attribute labels, and natural language descriptions [23].

D**. Other Datasets and Challenges:** The field of research constantly evolves, and new datasets may have been introduced in every fraction of time. Therefore, we discussed above some primary data set and also recommend using of best suitable dataset which will help to your research in the field of Multi-Modal Image Retrieval.

Various application of this field till now many challenges that need foster innovation in research. We also study to choose a useful technique for choosing the edge weights for the multi-layer graph such that the rankings for the images are optimized. Our basic analysis table demonstrate that the suggested approach outperforms state-of-the-art alternatives because it computes image similarity in a more meaningful way [24] .This paper's data collection is devoted to the pressing yet difficult issue of picture retrieval. By effectively mixing multimodal characteristics for picture ranking, it attempts to reduce the problems brought on by the so-called semantic gap. Currently, academics working on picture retrieval utilize quite complex algorithms to address this issue [24].

V. EVALUATION MATRICS FOR MULTIMODAL IMAGE RETRIEVAL

In Multi-Modal Image Retrieval research, evaluation metrics are essential for evaluating the effectiveness of algorithms and models. A model is mandatory for implementation or working for a dataset; So that metrics are play an analysis tool to check every phase of performance [25]. They measure how well the system can locate pertinent information using various modalities (such as text and graphics) [25]. These assessment measures for multi-modal image retrieval are frequently used.

(A)Precision and Recall :Precision and Recall are two fundamental evaluation metrics used in information retrieval and classification tasks to measure the performance of a system. In this paper we mainly focus about to image retrieval by cross modalities. When we technically secure the performance metrics these two measurement technique are used. Precision is a measure of the accuracy of positive predictions made by a system [26]. It calculates the proportion of true positive instances among all the instances predicted as positive. Other one is define as Recall, also known as sensitivity or true positive rate, measures the proportion of true positive instances that were correctly identified by the system. It answers the question: "Of all the

actual positive instances, how many did the system correctly identify as positive?"[26].

(B) ***Mean Average Precision (MAP)***

Mean Average Precision (mAP) is a commonly used evaluation metric in information retrieval and object detection tasks. It is particularly useful when dealing with tasks involving ranked lists of items, such as document retrieval, image retrieval, or object detection. Mean Average Precision (mAP) evaluates how accurately and effectively a retrieval system performs by balancing precision and recall at various retrieval ranks [27]. It calculates the average precision for each query and then takes the mean of those average precision scores to yield a single summary metric, offering a comprehensive measure of the system's overall performance[26][27]. Overall, mAP is a comprehensive and widely adopted metric that provides a clear and meaningful assessment of the performance of retrieval systems, especially when dealing with tasks involving ranked lists and varying levels of relevance in the dataset [25].

(C)Normalized Discounted Cumulative Gain (NDCG)

Normalized Discounted Cumulative Gain (NDCG) is a widely utilized evaluation metric in information retrieval and recommendation systems [28] [29]. Its purpose is to evaluate the performance of ranked lists by assessing how well the system ranks relevant items higher in the list, taking into account both the items' relevance and their positions. NDCG considers the graded relevance of items, which indicates how relevant an item is to a given query. It applies a discount to the relevance scores based on the item's position in the ranked list [29] [27]. The discounted cumulative gain at a specific rank is computed by summing the relevance scores of relevant items up to that rank, with a discount factor that emphasizes higher-ranked items. Subsequently, the ideal DCG (iDCG) is calculated, representing the highest achievable DCG at that rank if all the relevant items were ranked at the top[29]. The formula for calculating NDCG at rank k is as follows:

Discounted Cumulative Gain (DCG):
1-DCG@k = Σ(relevance_score_i / log2(i+1)), where i ranges from 1 to k and relevance_score_i is the graded relevance score of the item at position i.
2-Ideal DCG (iDCG): To compute iDCG@k, sort the relevant items in descending order based on their relevance scores and then calculate DCG@k for this sorted list.
3-Normalized DCG (NDCG):
NDCG@k= DCG@k / iDCG@k

Overall, NDCG is a powerful metric that provides a comprehensive assessment of the effectiveness of ranked retrieval and recommendation systems, considering both relevance and ranking position[28].

(D) ***Rank Correlation Measures***: Rank correlation measures are statistical tools utilized to gauge the resemblance or concordance between the rankings of two distinct datasets of items. In research, these measures find widespread application across multiple domains, such as information retrieval, machine learning, and data analysis, to evaluate the coherence or association between diverse ranking sequences. They prove particularly valuable when the precise values of the items hold little significance, and the emphasis lies on the relative arrangement of the items.

VI. CHALLENGES AND FUTURE DIRECTIONS

As the field of multi-modal image retrieval using deep learning continues to advance, several challenges remain to be addressed. These challenges stem from the complexity of handling heterogeneous data sources, learning effective joint representations, and scaling up to handle massive datasets. Additionally, as the technology evolves, there are several promising future directions that researchers can explore to overcome these challenges and further enhance the capabilities of image retrieval systems.

(A) Heterogeneous Data Fusion: One of the major challenges in multi-modal image retrieval is efficiently fusing information from diverse data modalities. Different modalities may have different characteristics and data distributions, making it challenging to find an optimal way to represent and combine them. Future research should focus on developing novel fusion techniques that can effectively capture the complementary information from different modalities and leverage this knowledge to improve retrieval performance.

(B) Cross-Model Learning: Multi-model image retrieval requires learning representations that can bridge the gap between different domains or modalities. As datasets from various domains become more abundant, there is a need to design models that can transfer knowledge effectively across these domains. This includes investigating domain adaptation methods, transfer learning techniques, and unsupervised learning approaches to enhance the generalization ability of cross-model retrieval systems.

(C) Scalability: The growth of multimedia content and the ever-expanding size of image databases pose significant scalability challenges. Efficient and scalable retrieval algorithms are Essential to handle large-scale datasets in real-time. Future directions may involve exploring distributed and parallel computing techniques to speed up retrieval processes and designing compact and efficient models suitable for resource- constrained environments.

(D)Fine-Grained Retrieval: Fine-grained image retrieval, where the task is to retrieve visually similar images with subtle differences, is an emerging area of interest. Traditional image retrieval approaches struggle to capture fine-grained visual details. Future research should focus on developing specialized architectures that can learn fine-grained feature representations and improve the accuracy of retrieval for visually similar images.

(E) Multi-Task Learning: Leveraging multi-task learning can be beneficial for addressing data scarcity issues in some modalities and improving generalization. Developing multi-task learning frameworks that jointly optimize retrieval performance across multiple modalities can lead to more robust and effective image retrieval systems. Interpretability and Explain ability: Deep learning models are often considered as black-boxes due to their complex architectures. Interpretability and explain ability are critical aspects, especially in real-world applications such as medical image retrieval or safety-critical scenarios. Future research should focus on designing models that can provide insights into their decision-making process and justify their retrieval results.

(F) Real-World Applications: The ultimate goal of image retrieval research is to impact real-world applications

positively. Future research should strive to bridge the gap between academic advancements and practical implementations by conducting studies and experiments in real- world settings. Collaborations with industry partners can help validate the effectiveness and robustness of multi-modal and cross-model image retrieval systems in real-world applications. Recent trends in multi- modal and cross-model image retrieval using deep learning have shown promising results, but several challenges and exciting future directions lie ahead. Addressing these challenges and exploring new research areas will contribute to the development of more sophisticated and effective image retrieval systems, empowering users to access and utilize multimedia content in a more efficient and meaningful way.

VI. Conclusion

In this review paper, provides a comprehensive review of recent trends and challenges in multi-modal image retrieval using deep learning techniques. We explored various approaches and architectures employed in multi-modal retrieval, along with popular datasets discussion and evaluation metrics. Despite the significant progress achieved, several challenges remain, calling for further research in modality alignment, large-scale retrieval, multi-domain retrieval, and ethical considerations. By addressing these challenges, multi-modal image retrieval has the potential to revolutionize image search and retrieval applications in various domains.

References

[1] P. Udaiyar, "Cross-modal data retrieval and generation using deep neural networks Cross-modal data retrieval and generation using deep neural networks," 2020.

[2] A. Latif *et al.*, "Content-based image retrieval and feature extraction: A comprehensive review," *Math. Probl. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/9658350.

[3] S. Mishra and M. Panda, "Medical image retrieval using self-organising map on texture features," *Futur. Comput. Informatics J.*, vol. 3, no. 2, pp. 359–370, 2018, doi: 10.1016/j.fcij.2018.10.006.

[4] M. John, T. J. Mathew, and V. R. Bindu, "A Multi-Modal Cbir Framework with Image Segregation Using Autoencoders and Deep Learning-Based Pseudo-Labeling," *SSRN Electron. J.*, vol. 00, pp. 1–14, 2022, doi: 10.2139/ssrn.4067284.

[5] M. John, T. J. Mathew, and V. R. Bindu, "A Multi-modal CBIR Framework with Image Segregation using Autoencoders and Deep Learning-based Pseudo-labeling," *Procedia Comput. Sci.*, vol. 218, pp. 718–731, 2023, doi: 10.1016/j.procs.2023.01.052.

[6] S. Bickel, B. Schleich, and S. Wartzack, "A Novel Shape Retrieval Method for 3D Mechanical Components Based on Object Projection, Pre- Trained Deep Learning Models and Autoencoder," *CAD Comput. Aided Des.*, vol. 154, p. 103417, 2023, doi: 10.1016/j.cad.2022.103417.

[7] W. Chen, W. Wang, L. Liu, and M. S. Lew, "New Ideas and Trends in Deep Multimodal Content Understanding: A Review," *Neurocomputing*, vol. 426, pp. 195–215, 2021, doi: 10.1016/j.neucom.2020.10.042.

[8] B. McCartney, B. Devereux, and J. Martinez-del-Rincon, "A zero-shot deep metric learning approach to Brain–Computer Interfaces for image retrieval," *Knowledge-Based Syst.*, vol. 246, p. 108556, 2022, doi: 10.1016/j.knosys.2022.108556.

[9] B. Clip, V. Case, C. T. Lu, and F. Church, "Larissa Djoufack Basso in partial fulfillment of the requirements for the degree of Computer Science and Applications CLIP-RS : A Cross-modal Remote Sensing Image Retrieval Based on CLIP , Northern Virginia Case Study Larissa Djoufack Basso Abstract," 2022.

[10] Y. Wei and B. Akinci, "Panorama-to-model registration through integration of image retrieval and semantic reprojection," *Autom. Constr.*, vol. 140, no. May, pp. 0–11, 2022, doi: 10.1016/j.autcon.2022.104356.

[11] M. Milanova, "Deep Learning-Based Multimodal Image Retrieval Combining Image and Text," no. April, 2023, doi: 10.1109/CSCI58124.2022.00314.

[12] C. Xu, Z. Yu, X. Shi, and F. Chen, "Adding visual attention into encoder-decoder model for multi-modal machine translation," *J. Eng. Res.*, vol. 11, no. 2, p. 100077, 2023, doi: 10.1016/j.jer.2023.100077.

[13] F. Yan, A. M. Iliyasu, Y. Guo, and H. Yang, "Flexible representation and manipulation of audio signals on quantum computers," *Theor. Comput. Sci.*, vol. 752, pp. 71–85, 2018, doi: 10.1016/j.tcs.2017.12.025.

[14] J. Wang, "QRDA: Quantum Representation of Digital Audio," *Int. J. Theor. Phys.*, 2015.

[15] Y. Cheng, X. Zhao, R. Cai, Z. Li, K. Huang, and Y. Rui, "Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition," no. July, 2016.

[16] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, 2022, doi: 10.1109/TPAMI.2021.3059968.

[17] T. R. Ornob, G. Roy, and E. Hassan, "Informatics in Medicine Unlocked CovidExpert : A Triplet Siamese Neural Network framework for the detection of COVID-19," *Informatics Med. Unlocked*, vol. 37, no. December 2022, p. 101156, 2023, doi: 10.1016/j.imu.2022.101156.

[18] T. R. Ornob, G. Roy, and E. Hassan, "CovidExpert: A Triplet Siamese Neural Network framework for the detection of COVID-19," *Informatics Med. Unlocked*, vol. 37, no. December 2022, p. 101156, 2023, doi: 10.1016/j.imu.2022.101156.

[19] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[20] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations," *Trans. Assoc. Comput. Linguist.*, vol. 2, no. 1, pp. 67–78, 2014, [Online]. Available: https://aclanthology.coli.uni- saarland.de/papers/Q14-1006/q14-1006%0Ahttp://aclweb.org/anthology/Q14-1006.

[21] R. Yu, F. Jin, Z. Qiao, Y. Yuan, and G. Wang, "Multi-scale image–text matching network for scene and spatio-temporal images," *Futur. Gener. Comput. Syst.*, vol. 142, pp. 292–300, 2023, doi: 10.1016/j.future.2023.01.004.

[22] B. Englert and S. Lam, "On the use of XML for port communications," *IFAC Proc. Vol.*, vol. 42, no. 15, pp. 50–57, 2009, doi: 10.3182/20090902-3- US-2007.0059.

[23] R. Khasanova, X. Dong, and P. Frossard, "Multi-modal image retrieval with random walk on multi-layer graphs," *Proc. - 2016 IEEE Int. Symp. Multimedia, ISM 2016*, pp. 1–6, 2017, doi: 10.1109/ISM.2016.13.

[24] Z. Yuan *et al.*, "MCRN: A Multi-source Cross-modal Retrieval Network for remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 115, no. September, p. 103071, 2022, doi: 10.1016/j.jag.2022.103071.

[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[26] P. R. Sanchez and H. Zhang, "Precision spraying using variable time delays and vision-based velocity estimation," *Smart Agric. Technol.*, vol. 5, no. May, p. 100253, 2023, doi: 10.1016/j.atech.2023.100253.

[27] G. Cao, A. Iosifidis, M. Gabbouj, V. Raghavan, and R. Gottumukkala, "Deep Multi-View Learning to Rank," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1426–1438, 2021, doi: 10.1109/TKDE.2019.2942590.

[28] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002, doi: 10.1145/582415.582418.