Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.754736

MULTIMODAL FEATURE FUSION FOR IMAGE RETRIEVALUSING DEEP LEARNING

Vani Golagana [1], Prof. S. Vizianandа Row[2], Prof. P. Srinivasa Rao[3]
[1,2,3] Department of CSSE, AUCE, Andhra University, Visakhapatnam, AP, India.
[1] Corresponding author: email: vani.srr22@gmail.com

ABSTRACT

Multimodal data analysis, essential for processing information from diverse modalities like text and images, playsa crucial role in applications involving both these elements.As sentiment analysis gains popularity and multimedia data becomes ubiquitous, the integration of images and text proves beneficial across various fields such as image retrieval, image captioning, sentiment analysis, and recommender systems. In this study, we apply multiple image search methods, focusing on both visual and textual aspects. The primary objective is to analyze features in texts and images for the retrieval of relevantimages. Our approach revolves around a tripartite strategy. Firstly, we use text input vectors to retrieve images from extensive databases. Secondly, we compare text input vectors with combined text-image vectors. Thirdly, we propose directlycomparing fused text and image input vectors of the given query input vectors with fused vectors in the database. This multifaceted approach enables us to explore the relationships between textual and visual elements comprehensively. Ourwork concentrates on two tasks: individual feature extraction using encoding techniques and fusion strategies to concatenateboth text and image vectors. Addressing the need to capture detailed information, we incorporate visual and semantic features into our work. Natural language processing (NLP)and convolutional neural networks (CNNs) are employed to extract features from text and image data, respectively. After feature extraction, the features from multimodalities are fused using concatenation methods in our proposed Holistic Fusion Retrieval (HFR) model. This fusion of features enhances the relevance of extracted images, providing a more comprehensive representation of the underlying data. Our model (HFR) excels over other methods in performance, achieving an impressive average accuracy of 93% across five different contexts. This underscores its effectiveness in diverse scenarios and showcases its superiority in comparison to existing approaches.

Keywords: multimodal, feature extraction, fusion techniques

I. INTRODUCTION

The recent extraordinary progress in deep learning has propelled Computer Vision (CV) and Natural Language Processing (NLP) to new heights, achieving remarkable advancementsin various complex tasks. In the realm of computer vision, substantial breakthroughs have been made in visual content classification [1], object detection [2], semantic segmentation [3], and more. These achievements stem from leveraging large annotated datasets or implementing self-supervision techniques [4] on extensive unlabeled data. Simultaneously, in the field of

NLP, there has been a surge of interest in addressing multiple tasks concurrently through unsupervised pretraining of language models [5],[6],[7],[8] using extensive unlabeled corpora. However, there is a notable and growing enthusiasm for tackling challenges that necessitate the integration of linguistic and visual information, bridging the gap between traditionally distinct domains of computer vision and NLP. In multimedia data analysis, the synergy between textual and visual information has been a driving force in enhancing various applications, including image retrieval [9]. The advent of sentiment analysis [10], [11], a natural language processing technique aimed at discerning emotions and attitudes ex- pressed in text, offers a new dimension to the intricate field of relevant image retrieval [12]. Sentiment analysis enables us to extract the semantic content of textual descriptions associated with images and the emotional tone embedded within them which helps in refining the relevance of image search results. By integrating text and image features [13], we can craft a more nuanced representation that captures the visual content and the emotional narrative of images to further improve sentiment analysis. Traditional image retrieval methods often fail to capture the nuanced emotional and semantic dimensions embedded within images. Similarly, sentiment analysis that heavily relies on textual data may miss out on the valuable visual context presented in images. By integrating both modalities, we stand to enrich sentiment analysis with a comprehensive understanding of both visual aesthetics and textual context [14], [15], [16].

Feature extraction is a cornerstone of multimodal analysis [17] serving as the foundation for deeper analysis and retrieval. NLP techniques are employed for textual data to distill semantic meaning and sentiment polarity. On the other hand, Convolutional Neural Networks (CNNs) are utilized to extract intricate visual features from im- ages. These techniques enable the translation of unstructured data into meaningful representations, facilitating subsequent fusion and analysis. The fusion of multimodal features constitutes a key highlight of this research. Various fusion techniques, ranging from early fusion to late fusion and attention mechanisms [18], [19], offer avenues for harmonizing text and image data.

In our work, we use conventional encoding techniques to extract features from text and convolutional encoder techniques for extracting features from images. An early fusion strategy is employed to combine information from different modalities. The amalgamation of these modalities bridges the semantic gap and elevates the precision and relevance of content retrieval. The main goal is to identify the most effective techniques for retrieving relevant images in real-world contexts. This involves surpassing the longstanding separation between conventional domains and promoting a more seamless integration of linguistic and visual information. We will organize the remaining paper as follows: Section III outlines the related work on multimodal analysis, Section IV presents the methodology for building an image retrieval system based on multimodal systems. Finally, the results and discussions are presented in Section V, and the conclusions and future scope are given in Section VI.

II. RELATED WORK

Multimodal systems have transformed sentiment analysis and recommendation systems by

combining different data sources and enhancing traditional approaches. Multimodal sentiment analysis detects emotions expressed through multiple channels, including facial expressions, images, and text. Itacknowledges that emotions can be conveyed both verbally and visually. Lots of studies have found that using different types of information together (like pictures, text, and numbers) [20] is better than using just one type. In real- world scenarios, situations frequently entail a combinationof diverse data types. For instance, one might encounter a dataset for predicting prices alongside another containing medical records. This diversity in data types underscores the complexity of real-world applications, requiring adaptable and integrative approaches to effectively handle and analyze disparate data sources. Faliang Huang et al. [21] discuss various approaches for combining textual, visual, and acousticmodalities for sentiment classification and present datasetsand evaluation metrics commonly used in the field. Jie Xuet al. [22] propose various fusion architectures to integrate information from different modalities to understand sentiment and emotions comprehensively. These architectures combine modalities such as text, image, audio, and video to capture complementary cues and enhance the overall sentiment analysis performance. Zadeh et al. [23] focus on an entity-sensitive attention and fusion network to effectively model the intra-modality and inter-modality interactions. Hongyu Zho et al.

[24] aim to comprehensively review recent research efforts related to multimodal recommendation systems. Their work outlines a clear pipeline covering commonly used techniques ateach step, classifies models based on methods employed, and provides a code framework for new researchers to understand principles. Stuart J Miller et al. [25] propose a method for integrating natural language understanding into image classification to enhance classification accuracy. In recent years, researchers and practitioners have made substantial progressin various domains by analyzing and understanding individualdata modalities, such as text, images, and audio. As Computer Vision and Natural Language Processing (NLP) gained momentum in recent technological developments, cross-modal network learning for image-text similarity assumes a crucial role in query-based image retrieval tasks, and the application of image-text semantic mining can be harnessed effectively using cross-modal networks. T Abdullah et al. [26] highlight the growing significance of image-text matching in bridging the gap between heterogeneous visual and textual data. The paper offers an overview of recent advancements in image- text matching, focusing on deep architectures.

III. METHODOLOGY

The methodology covers three different ways of methodsin retrieving relevant images and the evaluation process for multimodal data analysis. We build a model for retrieving relevant images from a multimodal input containing both an image and text component. In the proposed workflow of our research, we delineate two crucial phases: feature extraction and fusion logic, tailored for multimodal data analysis. The feature extraction phase encompasses the extraction of both text and image features. Subsequently, we employ fusion strategies that amalgamate these heterogeneous feature sets into a single unified representation. This fusion logic seeks to exploit the synergy between textual and visual modalities, enabling a seamless integration of information early in the pipeline. The resulting fused representation serves as a com-

prehensive descriptor, facilitating more robust and informed downstream analysis, such as image retrieval.

A. Feature Extraction

It is a pivotal step in machine learning, involving the transformation of raw data into meaningful features tailored for optimal model input. An overview of feature extraction techniques for text encoding and image encoding reveals a diverse set of methods employed to distill key information from textual and visual data, enhancing the effectiveness of subsequent analysis and machine learning models.

Text Encoding: Text data, being inherently unstructured, istypically converted into vector representations for use as inputin various machine learning models. Specifically focusing on text encoding techniques such as word embedding andcontextual word embedding techniques, that capture semantic information and relationships among words.

Word Embeddings: Juan Ramos et al. [27] describe the main recent strategies for building fixed-length, dense, and distributed representations for words, capturing semantic relationships between terms using techniques like Word2Vec and GloVe.

The **GloVe** (Global Vectors for word representation) [28] algorithm initiates its process by constructing a word-word co-occurrence matrix from a given corpus. This matrix captures the frequency with which words appear together within a fixed- sized context window. Subsequently, the algorithm initializes word vectors and bias terms for each word present in thevocabulary. The core of the GloVe algorithm lies in defining an objective function aimed at learning word vectors. These vectors, when their dot product is taken, should equal the logarithm of the probability of word co-occurrence. The iterative refinement process of GloVe involves adjusting the word vectors and biases to minimize the defined objective function. As a result of this iterative refinement, the word vectors generated by GloVe encapsulate semantic relationships and syntactic structures. The GloVe objective function shown in Equation (1)is designed to learn word embeddings that capture semantic relationships based on the global co-occurrence statistics of words in a corpus.

$$\boldsymbol{J = \sum_{i,j=1} f(X_{ij})(w_i^T \ \ \tilde{w}_j + b_i + \tilde{b}_j - \log (X_{ij}))^2} \qquad \textbf{(1)}$$

Where, J is the overall cost function and V is the size of the vocabulary. X_{ij} represents the word co-occurrence count for words i and j. $\mathbf{w}_i$ and $\mathbf{w}_j$ are the word vectors and b_i and b_j are bias terms. $f(X_{ij})$ is a weight function defined as,

$$f(X_{ij}) = \min\left(1, \left(\frac{X_{ij}}{X_{max}}\right)^{\alpha}\right)$$

where X_{max} is a chosen threshold and α is a hyperparameter.

Contextual Word Embeddings: The contextual word embeddings involve utilizing models like BERT (BidirectionalEncoder Representations from Transformers) and USE (Uni-versal Sentence Encoder) that can capture the meaning ofwords based on their context within a sentence or document.Pre-training BERT [29] captures word meaning in a context-sensitive manner. It aims to learn contextualizedrepresentations of words that capture their meaning in the context of the sentence or document. This bidirectional context and dual pre-training objectives empower BERT to captureintricate nuances of language, providing contextualized wordembeddings that outperform previous models on a myriad of downstream tasks.

The Universal Sentence Encoder (USE) model [25] utilizes a type of encoding known as sentence embedding or semantic encoding. Sentence embedding involves converting variable-length sequences of words (sentences or phrases) into fixed-size vectors while preserving semantic meaning. Thearchitecture of the USE model is based on deep learning and it employs a combination of recurrent and transformer neural network components. The model is trained on a diverserange of data to understand the semantic relationships betweenwords and phrases. Once trained, the USE model is capable ofproducing fixed-size vector representations (embeddings) for input sentences, capturing their semantic content.

In the process of text feature extraction, our primary emphasis is on enhancing the representation of textual contentfor downstream tasks, as depicted in Figure 1. Utilizing established pre-trained models like GloVe, USE, and BERT, we generate embeddings to capture the intricate semantic information embedded in the text. This contributes to elevating model performance, ensuring more accurate and refined resultsin subsequent tasks.

Image Encoding: Image encoding plays a pivotal role in computer vision tasks, converting raw visual data into meaningful numerical representations that can be efficiently processed by machine learning models. Convolutional Neural Networks (CNNs) utilize multiple layers of convolutional operations to extract hierarchical features from images. Prominent CNN architectures include VGG (Visual Geometry Group), ResNet (Residual Networks). Idesai et al. [30] compared three transfer learning models namely VGG16, ResNet50, and Xception to determine their effectiveness in image classification. They evaluated each model's accuracy in predicting the classification of images. VGG16 and Xception, known for simplicity, excel in image classification but canbe computationally expensive. ResNet50, with residual con-nections, achieves state-of-the-art performance and is efficientfor transfer learning. ResNet50 is highlighted for its favorable balance between accuracy and efficiency.

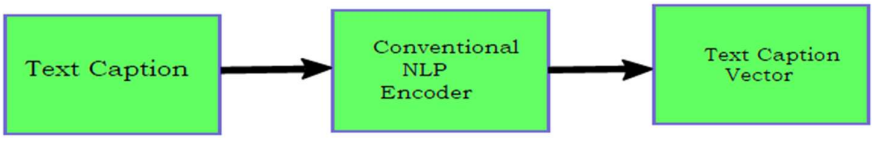


Fig. 1: Feature extraction from text caption

In image feature extraction, we prepare images appropriately for subsequent tasks using the

image preprocessing pipeline, and the resulting embeddings accurately capture the inherent features of the image shown in Fig 2.

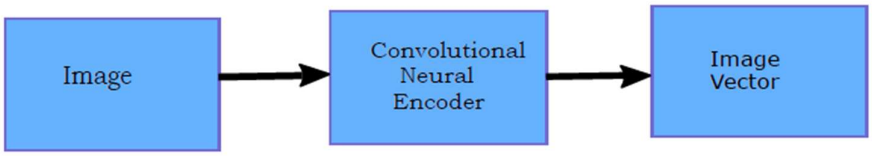


Fig. 2: Feature extraction from image

Deep convolutional neural network architecture [13] excels at automatically learning hierarchical features at different levels of abstraction. ResNet, or Residual Network, introduced by Microsoft Research in 2015, provides a solution to the vanishing/exploding gradient problem in deep neural networks. Utilizing skip connections in residual blocks, ResNet allows effective learning of residual mappings, enabling the training of highly deep networks. Initially based on a 34-layer plain network inspired by VGG19, the architecture is then trans- formed into a residual network with shortcut connections. ResNet variants with increased layers, such as 50, 101, or 152 layers, have demonstrated state-of-the-art performance in image recognition tasks.

In our work, we used ResNet50, a version of ResNet with 50 layers, each containing residual blocks with 3 convolutional layers. These blocks have convolutional layers, batch normalization, and ReLU activation function. The key to ResNet50 is the addition of shortcut connections (skip connections), which help the model learn residual mappings. This makes it easier to train deep neural networks, and ResNet50 is well-suited for our needs.

B. Data Fusion

Fusion techniques are crucial in various applications that involve multi-modal data integration and analysis, namely Early Fusion, Late Fusion, and Hybrid Fusion. Early Fusion is characterized by the amalgamation of raw or feature-level data from distinct sources or modalities at the inception of data processing, creating a unified representation for subsequent analysis. In contrast, late fusion maintains autonomous processing streams for each modality, deferring the information integration until the final decision-making phase, thereby enabling more independent modeling. A hybrid-based model strikes a balance between these two extremes, blending elements of early and late fusion to harmonize the capture of shared information while preserving modality-specific characteristics [31], [32], [33], [34]. The early fusion technique is a method employed in the field of data integration, particularly in scenarios where information from multiple modalities, such as text and images, needs to be combined for analysis or retrieval purposes. The numerical representations from both modalities are concatenated, meaning they are joined together to form a single, unified feature vector. This combined vector incorporates information from both textual and visual domains, creating a comprehensive representation of the input data. The resulting feature vector is then stored in

a database as shownin Figure 3.

The database is constructed to contain fused vector representations of both textual and image information sourced from the Flickr 8k dataset. The features for both modalities (Text and Image) are extracted using specialized techniques—CNN for images and NLP techniques for text. Subsequently, an early fusion technique is applied to merge thetwo sets of feature vectors, resulting in a unified representation. This fused feature vector database serves as a repository of integrated information, facilitating efficient retrieval and analysis.

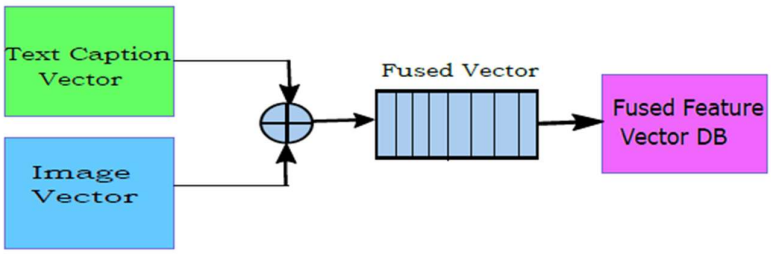


Fig. 3: Fusion of text and image vectors and the resultingDatabase

In our work, we have applied three different methods of retrieving relevant images based on text-centric and text-imagepaired data. These methods are outlined as follows:

Method-1: Text Semantic-based Retrieval (**TSemR**)

Method-2: Fused Semantic-based retrieval (**FSemR**)

Proposed Model: Holistic Fusion-based retrieval (**HFR**)

Method-1 Text Semantic-based Retrieval: In the TSemR approach, we establish a dedicated database that incorporates text vectors corresponding to textual data extracted from the Flickr 8k dataset, along with their respective image IDs. The textual data is transformed into vector representationsutilizing various models as discussed earlier in the featureextraction section. In the TSemR methodology, the primary focus is on the text query input. This input transforms into vector representations, employing the same techniques utilized during the initial construction of the database. The retrieval process unfolds by comparing these vectorized text inputs withthe corresponding text vectors stored in the database. This comparative analysis aims to identify and retrieve relevant images based on the semantic information encoded in thetext. The retrieval mechanism is visually depicted in Figure4, illustrating the process of matching text query with stored text vectors to retrieve pertinent images.

Fig. 4: Relevant images retrieved by TSemR

As a result, this procedure facilitates the retrieval of images whose associated vectors align with the input query's textvector, thereby delivering a refined selection of the top-k relevant images.

Method-2 Fused Semantic-based Retrieval: A feature vector for query text is constructed and is compared withthe fused input vectors stored in the fused feature vector database, which was presented in earlier section as shown in Figure. 3. Subsequently, based on the relevance determined through this comparison, relevant images are retrieved. This process is elucidated in Figure 5.

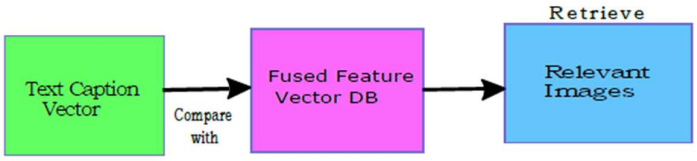


Fig. 5: Relevant images retrieved by FSemR

Proposed Model: In the proposed Holistic Fusion-based Retrieval (**HFR**) model, we present an innovative approachfor retrieving relevant images from a multimodal input that comprises both image and text components, as depicted in Figure 6. The process is encapsulated in Algorithm 1, outliningthe steps for feature extraction from both modalities and the subsequent retrieval of the top-k relevant images.

The input to this approach comprises both text and image data. Features from both modalities are extracted using the techniques discussed in the feature extraction section. Subsequently, the resulting feature vectors from both the text and im-age components are combined using an early fusion techniqueat the input level for further processing. This fusion process enables the integration of textual and visual information, enhancing the retrieval process by capturing comprehensive semantic representations from both modalities. This innovativemethod involves comparing the fused query vector with the vector in the fused feature vector database. The objective is to retrieve top-k relevant images based on the degree of relevance. The performance of the proposed model (HFR) vis-a-vis other methods namely TSemR and FSemR is presentedin the following section.

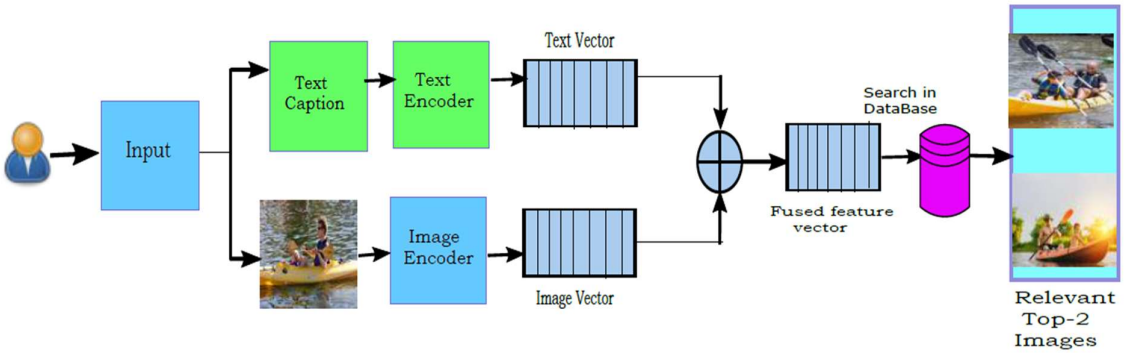


Fig. 6: Proposed Workflow for retrieving relevant images using HFR

Algorithm 1: Image Retrieval based on Fused Image and Text Vectors
Data: Dataset containing images and correspondingtext descriptions
Input: Query image and its caption
Result: Top−k Relevant images
Step 1: Extract feature vectors for images and text using a pre-trained model. Denote these vectors as I_i for image features and T_i for text features;
Step 2: Fuse image and text vectors to create fused vectors, F_i, for each image-text pair: $F_i \leftarrow [I_i, T_i]$;
Step 3: Extract feature vectors for the query image and its caption: I_{query} and T_{query}, respectively, using the same pre-trained models;
Step 4: Fuse the query vectors to obtain the fusedquery vector, $F_{query\ result}$;
Step 5: Compute the cosine similarity between $F_{query\ result}$ and all stored fused vectors F_i; **for** *each stored fused vector* F_i **do**
Step 6: Compute similarity score:
$\text{Similarity}(F_{query\ result}, F_i) \leftarrow \frac{F_{query\ result} \cdot F_i}{\|F_{query\ result}\| \cdot \|F_i\|}$
Step 7: Rank stored image-text pairs based on theirsimilarity scores in descending order;
Step 8: Retrieve the top-*k* pairs with the highestsimilarity scores as search results;

IV. RESULTS

In our research, we used the Flickr 8k dataset, which comprises 8,000 images, each accompanied by five captions. We employed image encoding techniques to transform the images into vector notation. These encoding methods are chosen to derive a condensed and meaningful representationof the visual content in vector form. The experiments areperformed on a desktop with a 2.3 GHz Intel core i7 processorwith 32 GB of RAM, and an Nvidia GeForce GTX GPU. The GPU acceleration allowed us to efficiently process the large dataset and train the encoders using Python 3.10.

This research investigates the performance evaluation of three image retrieval methods discussed in section IV, namely TSemR, FSemR, and HFR. These methods leverage specific encoding techniques, namely USE, BERT, and GloVe. The fusion of these text encoding techniques with image encoding using ResNet50 forms the basis for our evaluation, which spans across different contextual scenarios. Table I presents information on the encoding techniques applied in TSemR, FSemR, and HFR.

Method	**Encoding-1**	**Encoding-2**	**Encoding-3**
TSemR	USE	BERT	GloVe
FSemR	USE+ResNet50	BERT+ResNet50	GloVe+ResNet50
HFR	USE+ResNet50	BERT+ResNet50	GloVe+ResNet50

TABLE I: Encoding techniques used in TSemR, FSemR, and HFR

Context	**Example**
context-1	A Dog is running and playing in a grass
context-2	children playing in the park near the grass
context-3	Few people can be seen climbing a mountain
context-4	few men racing each other
context-5	children and dogs are playing near the water with adults

TABLE II: Example contexts from Flickr8k dataset

Conducting an experiment to evaluate the performance of three image retrieval methods on the entire dataset containing 8k images presented challenges. To overcome this, we have chosen a curated subset with 1300 images derived from the Flickr8k dataset. To ensure a comprehensive analysis, we extracted five distinct contexts from the subset shownin Table II and applied the said three different methods to retrieve relevant images for each context. Evaluation methods for image retrieval often focus on precision and recall at various levels of ranks to assess the performance of retrieval algorithms accurately [35], [36].

These metrics provide insights into the performance of asystem, particularly when dealing with a ranked list of items.

Precision@K: It measures the accuracy of a model's predictions within the top-k results. It is calculated as the ratio of the number of relevant items in the top-k results.

$$Precision@k = \frac{Number\ of\ relevant\ items\ in\ top-k}{k}$$

Recall@k: It is the ratio of the number of relevant items in the top-k to the total number of relevant items.

$$Recall@k = \frac{Number\ of\ relevant\ items\ in\ top-k}{Total\ number\ of\ relevant\ items}$$

Precision@k focuses on the accuracy of the top-k predictions, while Recall@k emphasizes the coverage of relevant items within the top-k results. The comparative analysis of precision and recall across the five contexts unveils notable trends and variations in the performance of the

	Method	Encoding-1 (%)	Encoding-2(%)	Encoding-3(%)
Context-1	TSemR	93	85	91
	FSemR	90	96	76
	HFR	94	96	96
Context-2	TSemR	80	71	39
	FSemR	82	52	44
	HFR	93	86	86
Context-3	TSemR	86	43	49
	FSemR	82	32	24
	HFR	93	87	85
Context-4	TSemR	60	29	33
	FSemR	69	29	25
	HFR	87	84	78
Context-5	TSemR	92	92	77
	FSemR	93	90	90
	HFR	98	97	97

methods.

Table III: Quantitative recall assessment across various contexts

Table III outlines the recall values for relevant retrieved images across diverse contexts, in respect of TSemR, FSemR, and HFR methods involving three said encoding techniques. Notably, HFR consistently outperforms its counterparts, show-casing superior recall values across all contexts.

In Context-1, HFR demonstrates commendable performancewith recalls of 94%, 96%, and 96% for Encoding-1,2 and3 respectively. This trend persists in Contexts 2, 3, 4, and5, where HFR consistently achieves remarkable recall values, particularly excelling with Encoding 1 and 2 techniques.

The evaluation of three methodologies (TSemR, FSemR, and HFR) across three different encoding techniques has recall of 82.2%, 64%, and 57.8% for Encoding-1, Encoding-2,and Encoding-3, respectively. FSemR exhibited average recallvalues of 83.2%, 59.8%, and 51.8% for the corresponding encodings.

Method	Encoding-1 (%)	Encoding-2(%)	Encoding-3(%)
TSemR	82.2	64	57.8
FSemR	83.2	59.8	51.8

HFR	93	90	88.4

TABLE IV: Average recall values across five contexts

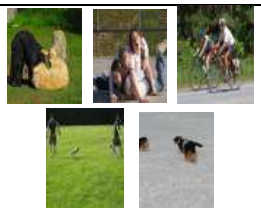
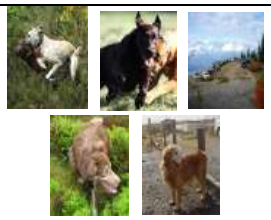
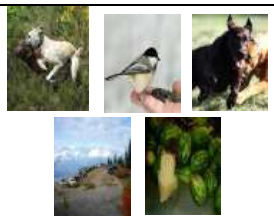
	Encoding-1(USE & ResNet)	**Encoding-2(BERT & ResNet)**	**Encoding-3(GloVe & ResNet)**
TSemR			
FSemR			
HFR			

Fig. 7: Image retrieval results for the caption ”Few men racing each other” across three methods for three encoding techniques

In contrast, HFR displayed superior performance withaverage recall rates of 93%, 90%, and 88.4% for Encoding-1, unveiled significant variations in average recall performanceacross all five contexts. TSemR demonstrated an average Encoding-2, and Encoding-3, respectively. These resultsemphasize the efficacy of the HFR methodology in consistently achieving higher recall rates compared toTSemR and FSemR across various encoding techniques.

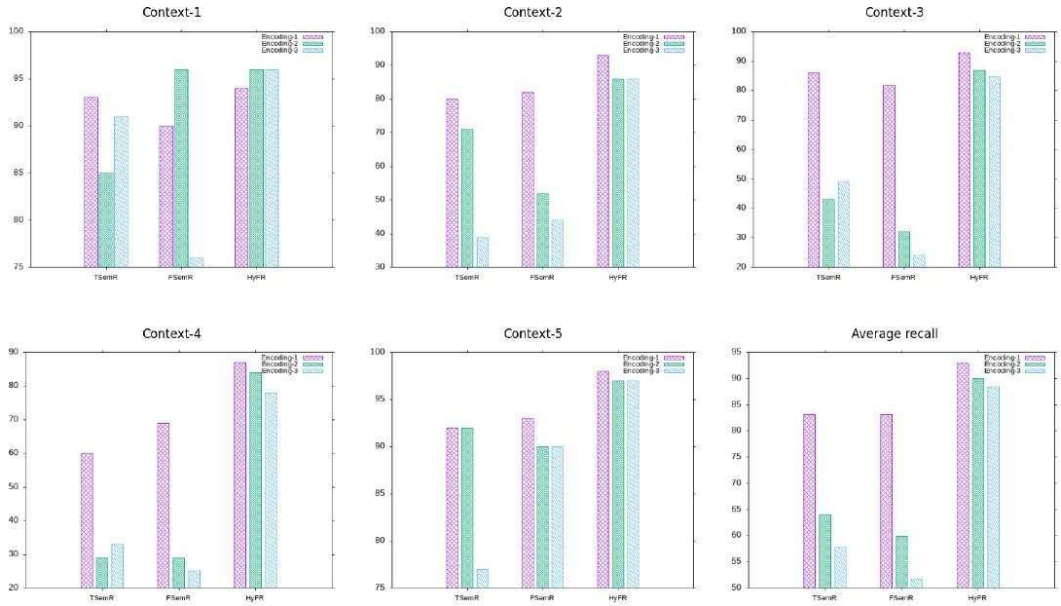


Fig. 9: Five plots showcasing recall values for different contexts, Accompanied by an aggregate plot for Average Recall

In the evaluation of three distinct encoding techniques, the retrieval outcomes for the given context, '**Few men racing each other**', reveal notable differences as shown in Figure.7 In case of encoding-1, both TSemR and FSemR identify three relevant images in their top 5 retrievals, while HFR outperforms by retrieving all five relevant images. In the case of encoding-2, TSemR and FSemR each yield one relevant image within the top 5, while HFR stands out once againwith a comprehensive set of five relevant results. Finally,in encoding-3, TSemR retrieves one relevant image, FSemR retrieves none, and HFR impressively retrieves all five rele- vant images within the top 5. Consequently, across all three encoding techniques, HFR consistently demonstrates superior performance in delivering precise and contextually fitting image retrievals for the specified query.

The plotting of recall values across the five distinct contextsshown in Figure 9, provides a comprehensive visual representation of the said three methods' performance in extracting relevant images through various encoding techniques.

After conducting a qualitative precision assessment across five diverse contexts, as shown in Table V, the average precision values for each method and encoding technique are shown in Table VI. Notably, HFR consistently outperformed both TSemR and FSemR in all encoding techniques and across different contexts. In Encoding-1, HFR demonstrated the high- est average precision of 24.6%, surpassing TSemR (21.8%) and FSemR (22.0%). Similarly, in Encoding-2 and Encoding- 3, HFR maintained its superior performance with average precision values of 23.6% and 23.4% respectively, compared to TSemR and FSemR. These results underscore the significance of HFR in consistently retrieving relevant information across various contexts.

	Method	Encoding-1 (%)	Encoding-2(%)	Encoding-3(%)
Context-1	TSemR	28	25	27
	FSemR	27	29	23
	HFR	28	29	29
Context-2	TSemR	20	17	10
	FSemR	20	13	11
	HFR	23	21	21
Context-3	TSemR	27	13	15
	FSemR	26	10	8
	HFR	29	27	27
Context-4	TSemR	17	8	9
	FSemR	19	8	7
	HFR	24	23	22
Context-5	TSemR	17	17	15
	FSemR	18	17	17
	HFR	19	18	18

Table V: Qualitative Precision Assessment for various Contexts

HFR consistently achieves higher recall rates compared to TSemR and FSemR across various encoding techniques, emphasizing its effectiveness in accurately retrieving relevant information. This superior performance underscores HFR'ssuitability for applications where recall is crucial. Its robust performance across diverse contexts and encoding methodologies positions HFR as a compelling choice among the evaluated methodologies, suggesting its potential as a tool for reliable information extraction.

Method	Encoding-1 (%)	Encoding-2(%)	Encoding-3(%)
TSemR	21.8	16	15.2
FSemR	22	15.4	13.2
HFR	24.6	23.6	23.4

TABLE VI: Average Precision values across five contexts

V. CONCLUSION & FUTURE WORK

In conclusion, our extensive analysis of three distinct approaches for retrieving relevant images through multimodal data has provided valuable insights. The text-centric based retrieval focused solely on text and established a strong foundation for text-based image retrieval. However, our HFR model, which incorporates both textual and visual techniques, proved to be the most impressive. By combining and integrating text and image data, we achieved exceptional and reliable results. This harmonious partnership between textual and visual information truly sets the standard for accuracyin retrieved images. In essence, our research emphasizes the significance of multimodal analysis and its potential to revolutionize image retrieval methods. Furthermore, the roleof multimodal data extends beyond image retrieval,

playinga pivotal role in applications such as sentiment analysis and recommender systems.

The fusion of textual and visual information proves par- ticularly beneficial in sentiment analysis, where understand- ing both linguistic and visual cues enhances the depth of emotion comprehension. Similarly, in recommender systems, multimodal data allows for a more nuanced understandingof user preferences by considering both explicit preferences expressed in text and implicit preferences conveyed through visual content. This highlights the versatility and impact of multimodal analysis across diverse applications.

REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[2] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern *recognition*, pages 7263–7271, 2017.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computervision*, pages 2961–2969, 2017.

[4] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language un- derstanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitasklearners. *OpenAI blog*, 1(8):9, 2019.

[7] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32,2019.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901,2020.

[9] Jianfei Yu, Jing Jiang, and Rui Xia. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439,2019.

[10] Ramandeep Kaur and Sandeep Kautish. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pages 1846–1870, 2022.

[11] Priyavrat, Nonita Sharma, and Geeta Sikka. Multimodal sentiment anal- ysis of social media data: a review. *Recent Innovations in Computing: Proceedings of ICRIC 2020*, pages 545–561, 2021.

[12] Xiaojun Xue, Chunxia Zhang, Zhendong Niu, and Xindong Wu. Multi- level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and*

Data Engineering, 35(5):5105–5118, 2022.

[13] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

[14] Tao Zhou, Jiuxin Cao, Xuelin Zhu, Bo Liu, and Shancang Li. Visual- textual sentiment analysis enhanced by hierarchical cross-modality in- teraction. *IEEE Systems Journal*, 15(3):4303–4314, 2020.

[15] Tadas Baltru s̆aitis, Chaitanya Ahuja, and Louis-Philippe Morency. Mul- timodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[16] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[17] Mahitha Potti, Avinash Chalumuri, Vani Golagani, and DND Harini. Transformer and deep cnn-based product recommendation system. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2023.

[18] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. Mul- timodal music information processing and retrieval: Survey and future challenges. In *2019 international workshop on multilayer music repre- sentation and processing (MMRP)*, pages 10–18. IEEE, 2019.

[19] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4):338–349, 2018.

[20] William C Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. Multi- modal classification: Current landscape, taxonomy and future directions. *ACM Computing Surveys*, 55(7):1–31, 2022.

[21] Faliang Huang, Xuelong Li, Changan Yuan, Shichao Zhang, Jilian Zhang, and Shaojie Qiao. Attention-emotion-enhanced convolutional lstm for sentiment analysis. *IEEE transactions on neural networks and learning systems*, 33(9):4332–4345, 2021.

[22] Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and S Yu Philip. Visual sentiment analysis with social relations-guided multiattention networks. *IEEE Transactions on Cybernetics*, 52(6):4472–4484, 2020.

[23] Yan Cheng, Leibo Yao, Guoxiong Xiang, Guanghe Zhang, Tianwei Tang, and Linhui Zhong. Text sentiment orientation analysis based on multi-channel cnn and bidirectional gru with attention mechanism. *IEEE Access*, 8:134964–134975, 2020.

[24] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*, 2023.

[25] Stuart J Miller, Justin Howard, Paul Adams, Mel Schwan, and Robert Slater. Multi-modal classification using images and text. *SMU Data Science Review*, 3(3):6, 2020.

[26] Taghreed Abdullah and Lalitha Rangarajan. Image-text matching: Methods and challenges. *Inventive Systems and Control: Proceedings of ICISC 2021*, pages 213–222, 2021.

[27] Felipe Almeida and Geraldo Xex é o. Word embeddings: A survey, 2023.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 con- ference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language un- derstanding. *arXiv preprint arXiv:1810.04805*, 2018.

[30] Anil B Desai, DR Gangodkar, Bhaskar Pant, and Kumud Pant. Com- parative analysis using transfer learning models vgg16, resnet 50 and xception to predict pneumonia. In *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pages 1–6. IEEE, 2022.

[31] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026, 2020.

[32] Meng Xu, Feifei Liang, Xiangyi Su, and Cheng Fang. Cmjrt: Cross- modal joint representation transformer for multimodal sentiment analy- sis. *IEEE Access*, 10:131671–131679, 2022.

[33] Jianfei Yu, Jing Jiang, and Rui Xia. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439, 2019.

[34] Jiaxuan He and Haifeng Hu. Mf-bert: Multimodal fusion in pre-trained bert for sentiment analysis. *IEEE Signal Processing Letters*, 29:454–458, 2021.

[35] Venkat N Gudivada, Dhana L Rao, and Amogh R Gudivada. Information retrieval: concepts, models, and systems. In *Handbook of statistics*, volume 38, pages 331–401. Elsevier, 2018.

[36] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014.