

CONTEXT-AWARE IMAGE RETRIEVAL: ENHANCING SEARCH PRECISION IN LARGE-SCALE IMAGE DATABASES USING BLIP AND AUTOMATED CAPTIONING

JAHNAVI SOMAVARAPU¹, [0000-0003-0111-3392], RAVI KANTH MOTUPALLI¹, ANJANEYULU NELLURU¹, VENKATESWARA RAO KOTA², SUDHAKAR YADAV NALADESI³, TEJASWI POTLURI¹

¹Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology, Hyderabad, Telangana, 500118, India

²Andhra Loyola Institute of Engineering and Technology, Jayaprakash Nagar, Vijayawada, Andhra Pradesh 520008, India

³Chaitanya Bharathi Institute of Technology, Hyderabad, Osman Sagar Rd, Kokapet, Gandipet, Hyderabad, Telangana 500075, India

E-mail: ¹jahnavi_s@vnrvjiet.in, ²ravikanth_m@vnrvjiet.in, ³anjaneyulu_n@vnrvjiet.in, ⁴siva278@gmail.com, ⁵sudhakaryadavn_it@cbit.ac.in, ⁶tejaswi_p@vnrvjiet.in

ABSTRACT

The growing number of digital images has rendered efficient retrieval a serious issue, since conventional approaches based on low-level features like color and texture are not able to capture semantic content. Users often find it difficult to retrieve appropriate images because there is no context-aware search mechanism, particularly when precise metadata or visual information is not known. Current retrieval systems are unable to fill the semantic gap between text-based queries and image content, producing inaccurate or incomplete results. In order to solve this, we introduce a context-aware image retrieval system that combines deep learning and natural language processing (NLP) approaches. The system utilizes BLIP model for automatic image captioning, constructing a meaningful textual description of images. User requests are processed based on Sentence-BERT (SBERT) embeddings, coupled with TF-IDF and Levenshtein distance-based string matching to enable accurate retrieval. An intuitive user-friendly search is offered through a Gradio-based interface. Through this work, it is illustrated that visual feature extraction combined with NLP captioning and query processing boosts the accuracy of image search substantially. The system suggested here not only enhances the precision in retrieval but also maintains scalability and adaptability for big unstructured data. Future efforts will be dedicated to achieving further computational efficiency, including multimodal retrieval, and improving user personalization to improve the system responsiveness to various user needs.

Keywords: *Image Retrieval, Semantic Search, Deep Learning, Sentence-BERT, Query Encoding, Context-Aware Search, Gradio Interface.*

1 INTRODUCTION

With the rise of digital technology, the tremendous growth of image data has imposed unprecedented challenges on image organization, image search, and image retrieval effectiveness. Traditional image search tools using metadata, file names, or hand-tagging are most often unable to meet expectations in managing large collections of unstructured images. The users will be unable to search for popular images if they have difficulty remembering exact details like file names or dates. In an attempt to tackle this challenge, context-aware image retrieval systems came on board as a robust remedy owing

to the supremacy of deep learning, computer vision, and natural language processing (NLP) to support more accurate searching and better user satisfaction. This project is focused on developing an intelligent image retrieval system where image search is done based on descriptive natural language queries rather than keyword-based searching.

The system employs the BLIP model for automatic captioning of images, building a descriptive text of an image. To find similarities with queries, the system employs a new similarity model for NLP, Sentence-BERT (SBERT), to transform both user queries and captions into semantically rich vector representations. Through cosine similarity-based

calculation between query embeddings and caption embeddings, the system searches and scores images based on semantic similarity. Along with this, TF-IDF similarity and Levenshtein distance are applied to refine the outcome further despite varying phrasing within the query or minor spelling mistakes. The goal of this work is to recommend an easy and friendly image retrieval system that spans the gap between human language and machine-based image indexing. This method enhances precision in retrieval, minimizes human tagging dependency, and facilitates usability in practical applications like digital asset management, content-based image retrieval, and multimedia information systems.

Through the integration of deep learning-based visual feature extraction, NLP-based semantic perception, and multi-layered similarity matching, this work shows how effective it is to have a new, context-aware image retrieval system for large-scale and unstructured image databases.

The later sections in this paper cover, related works in section II, the details of the proposed system mentioned in section III, the results given in section IV, and the conclusion and future scope in section V.

2 RELATED WORKS

De Boer, M. H. T., Laura Daniele, Paul Brandt, and Maya Sappelli (2015) [1] explored the application of semantic reasoning techniques to enhance image retrieval accuracy. By establishing images with semantic descriptions and contextual data, the proposed technique supports more effective correspondence between user searches and retrieved pictures. The focus of this paper is on highlighting the importance of deep models incorporated with structured reasoning processes, achieving better search relevance and retrieval correctness. This work forms the foundation of modern context-aware retrieval systems designed to narrow the semantic gap in large image repositories.

Farouk and Mamdouh (2020) [2] introduced a discourse-based similarity model that quantifies sentence relations from composed linguistic features. This is a method that provides greater insight into the semantics of sentences, and therefore it is very beneficial in text-based image retrieval tasks. Through the use of discourse structures, the model effectively maps user

queries and image captions for better retrieval accuracy. The work describes how advanced sentence similarity techniques can enhance the readability of text queries, thereby optimizing the overall performance of context-sensitive image retrieval systems.

Zhang et al. proposed spatial-context-aware global features, and Yang et al. (2023) [3] employed attention mechanisms to achieve multi-scale fusion. Such research demonstrates that using structural relationships and semantic reasoning rather than vision similarity enhances the robustness and accuracy of retrieval. Our research further embraces these tenets by blending BLIP-generated captions with a hybrid matching strategy to enable precise alignment between image content and user intent.

Zhang, Qi, et al. (2020) [4] pioneered this approach with their Context-Aware Attention Network (CAAN), which dynamically integrates inter-modal region-word matching and intra-modal semantic relations. Building on this, BERT-optimized models like UNITER and ViLBERT achieved better performance with large-scale pretraining but incurred higher computational costs. These experiments collectively show that strong retrieval requires both precise visual-textual alignment and good semantic reasoning—principles underlying our light-weighted fusion of BLIP captions and spatial-context-aware features.

Wei, Wenzhang, et al. (2025) [5] introduced a sophisticated context-aware image retrieval system that utilizes deep learning and semantic reasoning to improve retrieval precision. Their research combines vision-language models to produce textual descriptions of images, enhancing search relevance by closing the semantic gap between visual content and user queries. The research highlights the significance of transformer-based models for encoding image features and text descriptions to enable more accurate multi-modal retrieval. Experimental results show better performance compared to conventional CNN-based methods, reflecting the efficacy of integrating natural language processing (NLP) methods with visual inspection. This work sets a solid foundation for smart image search by illustrating how semantic encoding and deep feature extraction enhance the accuracy of searches and the overall user experience on large image repositories.

Chen et al. (2022) [6] examined the improvement of semantic similarity modeling in text-based

retrieval through context-aware embeddings. They introduced a transformer-based model that boosts sentence representations to improve retrieval precision by aligning semantic structures with queries from users. The technique achieved significant improvements over word vector models by incorporating discourse-level reasoning, thereby enhancing the relevance of the retrieved text information. By addressing issues of sentence disambiguation and contextual alignment, this work provides a solid foundation upon which more effective natural language processing (NLP)-based query retrieval systems can be constructed. Study findings underscore the practicality of deep contextual embeddings, consistent with prevalent multi-modal retrieval models integrating image and text understanding.

Farouk (2019) [7] surveyed sentence similarity approaches, categorizing them into word-based (lexical similarity), structure-based (syntactic patterns), and vector-based (deep learning) approaches. The study showed that the combination of these approaches by using hybrid approaches that combine these approaches is best because they encode both structural and semantic relationships. These findings are in line with modern NLP systems that require robust similarity estimation for question answering and retrieval tasks.

Ge et al. (2021) [8] introduced SMFEA, a structured multi-modal feature embedding and alignment model for image-sentence retrieval to solve the problem of semantic and structural consistency across visual and textual modalities. The approach presents Visual and Textual Context-aware Structured Trees (VCS-Tree and TCS-Tree) to encode intra-modal relations and align inter-modal fragments explicitly via a shared referral tree. By utilizing KL-divergence for node-level matching and blending in-stance-level, structured, and consensus-aware characteristics, SMFEA surpassed current state-of-the-art techniques on Flickr30K and MS-COCO datasets with cross-modal retrieval robustness. The work emphasizes the need for explicit structural matching to narrow the heterogeneous gap between text and images.

Cui et al. (2024) [9] introduced a context-aware Relation Enhancement and Similarity Reasoning (RESR) model for image-text retrieval, which overcomes the shortcomings of previous methods by combining intra-modal relation enhancement and inter-modal similarity reasoning. The model utilizes a new Context-aware Graph Convolutional

Network (C-GCN) to boost local feature representations with semantic relations and global-context information, followed by a two-stream similarity reasoning strategy to improve cross-modal alignment. The experimental results demonstrate the ability of the model to process complex scenes and diverse queries, and it is a crucial advancement in vision-language combination.

Wu et al. (2025) [10] proposed the Syntactic-Guided Optimization of Image-Text Matching (SGIM) model, which optimizes redundancy in features during intra-modal image modeling and improves syntactic comprehension during text representations. The method uses multi-view filtering to control attention weights in image areas, removing redundant information and highlighting important areas. For text, it uses syntactic dependency parsing to build ancestor-descendant relationships, improving the model's capacity to learn long-range contextual dependencies. The paper emphasizes intra-modal feature refinement and syntactic structure fusion as critical to successful cross-modal retrieval, proposing a scalable approach towards multimedia analysis and semantic search applications.

Zhong Ji et al. (2024) [11] introduced the Hierarchical Matching and Reasoning Network (HMRN) to enhance multi-query image retrieval based on employing a structured hierarchical framework. Their method involves three innovations: Scalar-based Matching (SM) which applies symmetric cross-attention to compute inter-level similarities between image regions and text queries using self-attention mechanisms for gathering global context in the form of summed global contextual information; Vector-based Reasoning (VR), which constructs an inter-correlation graph and extracts high-level semantic correlations among many queries in the form of node updates with learnable affinity matrices; and hierarchical similarity aggregation strategy, which adds up multi-level features by weighted fusion with $\alpha = 0.4$ for local, $\beta = 0.4$ for reasoning, and 0.2 for global. Our research provides insightful guidance on structured multi-query retrieval and is complementary to our own research on context-aware image retrieval through improved semantic matching between user queries and the retrieved images in large-scale databases.

Baldrati et al. (2023) [12] proposed SEARLE, a zero-shot composed image retrieval (ZS-CIR) method that seeks to break dependency on labeled training data by using the vision-language space of

CLIP. Their approach translates reference images into pseudo-word tokens through textual inversion, which are subsequently combined with relative captions for retrieval. The major contributions of this paper are an optimization-based textual inversion (OTI) process, which is augmented with GPT-driven regularization to guarantee alignment with CLIP's token manifold, and a light-weight multi-layer perceptron (MLP) distilled from OTI for efficient inference. SEARLE surpasses current supervised models on FashionIQ, CIRRR, and the newly proposed CIRCO dataset, which involves multiple ground truths for better evaluation. By taking advantage of unsupervised training on unlabeled images and the addition of a hybrid semantic-textual fusion mechanism, this method closely follows our project's goal of improving image retrieval efficiency using multi-modal embeddings and is thus a suitable basis for enhancing query-based retrieval accuracy.

Dong et al. (2023) [13] proposed the Region-to-Patch Framework (RPF) for attribute-specific fashion retrieval, with both CNN-based region-aware processing and ViT-based patch-aware refinement to achieve higher precision for subtle attribute matching. Their framework utilizes a region-aware branch for coarse localization of attribute-related areas, e.g., sleeve styles, and uses a patch-aware branch to harvest fine-grained details for higher retrieval accuracy. The multi-scale feature fusion and contrastive learning methods investigated in this research work are of great significance for improving attribute-aware retrieval systems, validating the potential of using hierarchical feature extraction for more accurate image search.

Tang et al. (2024) [14] introduced Context-I2W, a zero-shot composed image retrieval (ZS-CIR) method that dynamically maps reference images onto context-dependent pseudo-word tokens with manipulation descriptions as contextual inputs. Compared to previous techniques that project images into static pseudo-words, Context-I2W consists of two key components: an Intent View Selector, which aligns image embeddings with task-specific intents such as domain conversion and attribute editing, and a Visual Target Extractor, which employs learnable queries to extract fine-grained local information. The ability to generalize efficiently without task-specific supervision highlights the flexibility of the framework, making it a valuable addition to the creation of zero-shot image retrieval techniques.

Seonwoo et al. (2022) [15] proposed RankEncoder, an unsupervised sentence representation learning approach that improves sentence embeddings by injecting neighborhood relations from an external corpus. Conventional sentence encoders mostly depend on standalone sentence representations without considering the global semantic context provided by similar sentences. RankEncoder addresses this drawback by producing rank vectors, normalized rankings of all corpus sentences by their similarity to a target input. Although the technique is mostly tailored for monolingual text processing, its emphasis on using corpus-wide semantics holds promise for application to our project, specifically in improving caption coherence and boosting text-based similarity retrieval in a cross-modal retrieval framework.

Li, Jinhang, and Yingna Li (2022) [16] introduced a sentence-matching model with multi-granularity representations to boost text similarity assessment accuracy. The technique combines token-level fine-grained interactions and coarse-grained sentence embeddings by making use of hierarchical attention for richer semantic comprehension among sentence pairs. The model introduces a multi-perspective matching layer to learn multiple kinds of similarity patterns, a bidirectional interaction module to postprocess sentence alignment, and an adaptive fusion approach to dynamically integrate local and global textual dependencies. The work provides beneficial insights on better semantic representation learning, aligning with the purposes of optimizing text-based retrieval systems in various natural language processing tasks.

Gardazi et al. (2025) [17] thoroughly analyzes BERT's revolutionary effect on NLP tasks such as sentence boundary detection, tokenization, named entity recognition (NER), and sentiment analysis. The authors note BERT's bidirectional contextual embeddings, which surpass conventional models such as Word2Vec and LSTM by picking up subtle semantic relationships. Main adaptations like ALBERT, Distil-BERT, and RoBERTa solve computational issues while keeping performance intact. This paper fits the current requirements in context-aware systems, providing knowledge on how to use BERT's architecture in applications such as image retrieval, where semantic meaning fills the gap between user search and visual data.

Li et al. (2024) [18] presented the Comateformer, a novel transformer-based semantic sentence matching model that addresses the limitations of typical attention mechanisms in detecting fine-grained differences between sentence pairs. The model incorporates an attention aggregation mechanism that replaces the traditional softmax with a dual-affinity module to capture similarity and dissimilarity among words simultaneously. The method allows the model to perform scaling, subtractive, or additive operations on representations of tokens that can lead to more subtle differences between semantically close but disparate sentences. That the model achieves this shows great promise for using compositional attention to enhance pre-trained language models in practical uses.

Li, Wei, et al. (2024) [19] introduced CAT-LLM, a context-aware training-augmented large language model specially designed for multi-modal contextual image retrieval (MMCIR). It addresses the limitations of existing methods by leveraging LLMs to better understand complex multi-modal inputs, such as pairs of texts and images. The model is conditioned on two task-specific objectives: Context-Aware Captioning (CA-Cap) and Context-Aware Text Matching (CA-TM), which enhance its ability to integrate visual and text information. Its performance is further confirmed by ablation studies and qualitative analysis, and it is found to be beneficial in processing challenging and varied retrieval tasks.

Li, Bo, Di Liang, and Zixin Zhang (2024) [20] proposed PlugIR, a plug-and-play interactive text-to-image retrieval system based on large language models (LLMs) to enhance retrieval performance through dialogue-based interactions. The system maps dialogue-form queries into query-friendly formats of pre-trained retrieval models without fine-tuning. It also employs an LLM questioner to generate non-redundant, context-aware questions from retrieval candidates to improve efficiency and accuracy. PlugIR offers the Best log Rank Integral (BRI) metric for robust evaluation, outperforming zero-shot and fine-tuned baselines on numerous benchmarks. Generalizability to various retrieval systems, including black-box systems, and insensitivity to context perturbations demonstrate its robustness and applicability in real-world scenarios.

3 METHODOLOGY

The context-aware image retrieval system enables users to search for images using phrases in a quick and simple method. A combination of deep learning, natural language processing (NLP), and similarity computation techniques is used to relate user queries with appropriate images in an extensive unorganized image database. The logic of the system is divided into several stages, allowing for precise and contextualized image retrieval.

(i) Image Processing and Feature Extraction

The processing and extraction of images is critical to the system because it transforms unprocessed pictures into high level structured forms that can be related to the user's requests. Traditionally Convolutional Neural Networks (CNNs) have been used to solve this problem, however, the proposed system implements a Vision Transformer (ViT) to improve the performance as shown in figure 1. Instead of CNNs, which employ local receptive fields and hierarchically organized feature maps, ViTs treat images as a collection of non-overlapping patches and use global self-attention to summarize context. This allows the model to better capture intricate patterns, textures, and relationships between objects.

The features are treated as numerical vectors which represent the most relevant parts of an image and can be easily compared with others. To facilitate quick retrieval, these vectors are kept in a well-organized database with other vectors alongside the textual information. The choice of ViT over CNNs was made because it greatly outperforms in dealing with various types of images and gives richer embeddings that correspond to the users' descriptions of the queries, thus providing accurate matches.

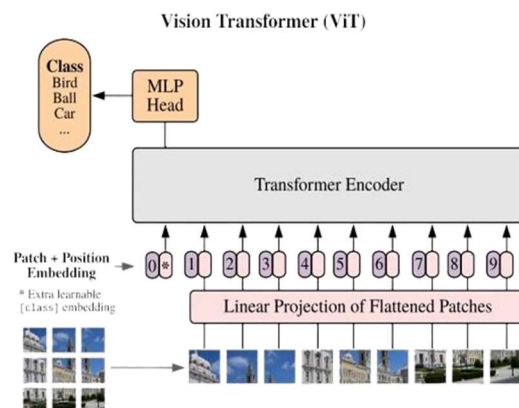


Figure 1: Vision Transformer

(ii) Image captioning

To improve the retrieval procedure, images are provided captions through the use of BLIP (Bootstrapped Language-Image Pretraining). BLIP is an integrated advanced vision-language model that utilizes a Vision-Transformer (ViT) image encoder and a text generating transformer to produce a pretative text words regarding the image. This model is capable of image understanding and captioning it due to having been trained on large scale image caption pair datasets.

In the BLIP processing, image captioning involves the identification of potential captions that BLIP associates with the image content. Captions do facilitate retrieval but also further enrich the image database semantically. BLIP, in contrast to traditional approach of object-detection, goes further by capturing the context such as the relationships between objects, what actions are taking place, and the environment in which the action ae occurring. The ability to search using natural language queries that more accurately coincide with the image content helps improve search accuracy.

Along with a set of feature vector, captions created throught the process are saved in the database. This method gives two formen of image retrievals, one where images are accessed through visual embeddings and the second where images are retrieved through textual descriptions.

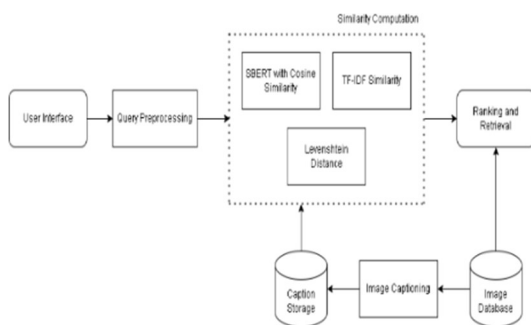


Figure 2: System Architecture

(iii) Text Processing and Query Encoding

The performance of the image retrieval system depends on its ability to interpret and process queries effectively. To achieve this, the system uses Sentence-BERT (SBERT) (coupled with cosine similarity), which is a variation of BERT specially designed for sentence-level semantic similarity tasks. SBERT converts text descrip-

tions into dense vector representations that capture meaning as well as simple word matching. These query embeddings are then compared against precomputed image caption embeddings to identify the best matches.

In addition to SBERT, the system also employs TF-IDF (Term Frequency-Inverse Document Frequency) to enhance keyword-based matching further. TF-IDF helps identify the most important words in a query to improve the system's ability to prefer important words. The system also utilizes Levenshtein Distance, a string-matching metric that estimates the number of character changes needed to transform one string to another, to facilitate variation in user input, such that small spelling mistakes or reworded queries do not negatively impact retrieval performance. The fusion of SBERT, TF-IDF, and Levenshtein Distance results in a solid text-processing pipeline that substantially enhances the accuracy of semantic-based image retrieval.

(iv) Image Retrieval and Ranking

After the user provides a search query, the system follows a multi-layered retrieval and ranking operation to provide the most relevant images. The cosine similarity measure is applied to measure the similarity between the SBERT-produced query embedding and caption embeddings stored elsewhere to determine their similarity. Because SBERT encodes the meaning of sentences instead of words, the process facilitates very accurate image retrieval based on contextual meaning.

In conjunction, TF-IDF scores are calculated to evaluate how well the query is aligned with single words in the image captions as a secondary ranking factor. Levenshtein similarity scores are also calculated to capture slight text differences. A weighted sum of the three scores (SBERT cosine similarity, TF-IDF similarity, and Levenshtein similarity) is utilized to establish the final image ranking. This approach ensures that images more semantically and contextually close to the query are placed at the top of results.

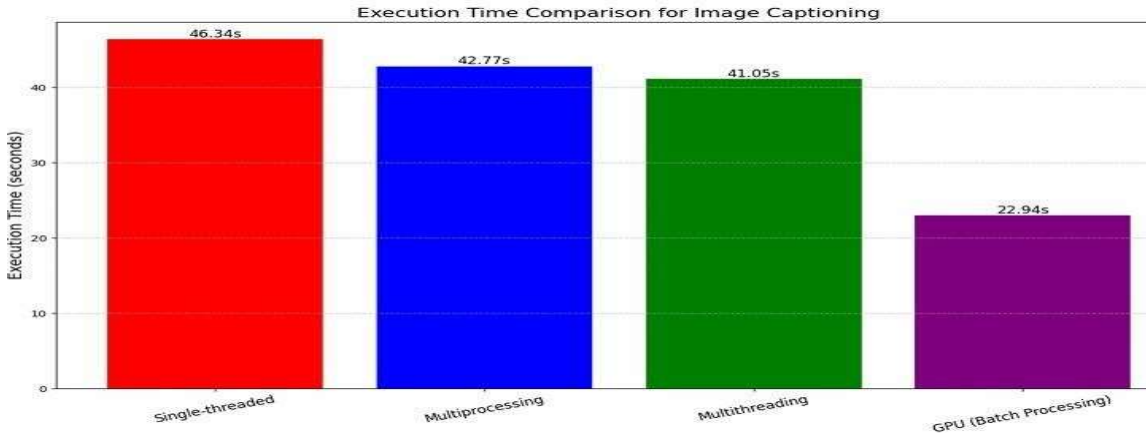


Figure 4.2.1: Execution time variance with various methods

4.3 Search Efficiency and Relevance

The performance of the voice and text-based image retrieval system was tested with semantic search queries as shown in Figure 4.3.1. The system was able to retrieve and rank appropriate images from textual descriptions, exhibiting strong natural

language understanding. Figure 4 also illustrates the top 5 search results by relevance scores, where the most precise match achieved the highest ranking. The system is able to effectively differentiate between close captions, achieving high-precision retrieval.

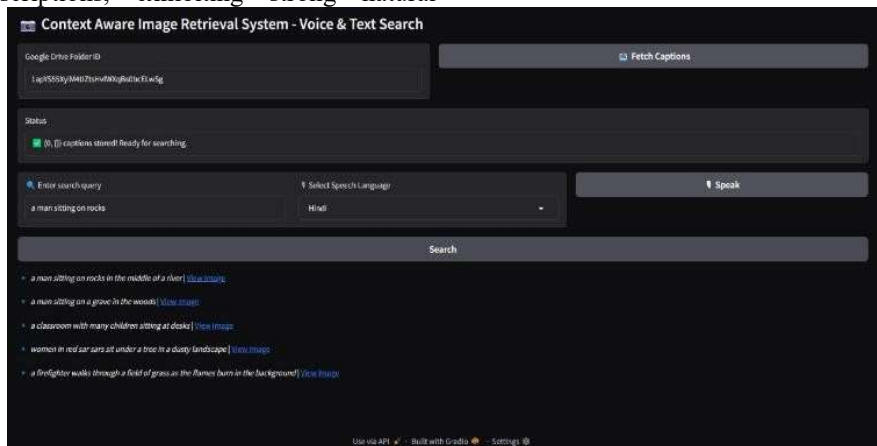


Figure 4.3.1: User Interface showing results for the query

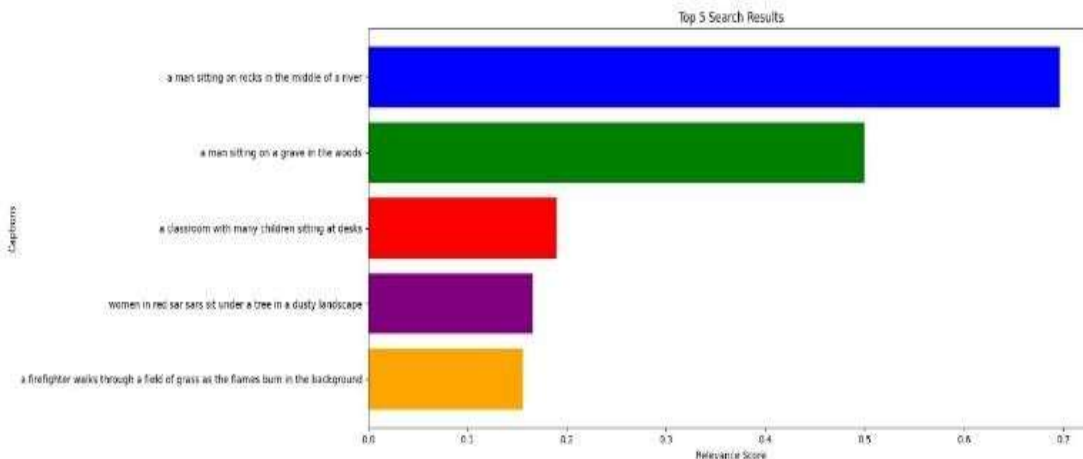


Figure 4.3.2: Similarity of Result images with the query

5 CONCLUSION AND FUTURE SCOPE

This paper introduces a context-aware image search system that improves search precision in large, unstructured image databases. Through the use of Convolutional Neural Networks (CNNs) for feature extraction and BLIP for automated captioning, the system accurately pairs textual query with corresponding images. The use of deep learning and Natural Language Processing (NLP) dramatically enhances the accuracy of retrieval, enabling searches to be more intuitive and efficient. Experimental outcomes verify that the system proves effective in facilitating the semantic bridge between user search and image meaning, surpassing conventional image searching methods. Even though the system performs exceptionally in retrieving semantically and visually aligned images, improving the quality of caption generation presents some difficulties with very abstract search queries.

Future research can focus on developing query-aware or knowledge-enhanced captioning models that incorporate external semantic knowledge and common sense reasoning to better capture abstract meanings. Additionally, integrating hybrid retrieval frameworks that combine caption-based indexing with multimodal embedding similarity may further improve retrieval performance for abstract and concept-driven queries. Further development can improve the performance of the system by using sophisticated NLP models such as transformers for improved query comprehension. Multimodal retrieval, with the inclusion of metadata like geolocation and timestamps, can further improve search results. Scaling the system to handle real-time, large-scale data using cloud computing and optimized indexing is another major area. Other enhancements include multilingual and voice-based search integration, allowing for greater accessibility. Personalization via adaptive learning and user feedback might shape search rankings more and more over time. Furthermore, expanding the system to domain-specific uses, like medical imaging or e-commerce, would make it even more useful in the real world. Finally, integration with AR/VR could add interactive search experiences.

In summary, this study presents an intelligent and scalable image retrieval system, with future development in deep learning and NLP ensuring even higher efficiency and flexibility across different fields.

REFERENCES:

- [1] De Boer, M. H. T., Laura Daniele, Paul Brandt, and Maya Sappelli. "Applying semantic reasoning in image retrieval." (2015).
- [2] Farouk, Mamdouh. "Measuring sentences similarity based on discourse representation structure." *Computing and Informatics* 39.3 (2020): 464-480.
- [3] Zhang, Zhongyan, et al. "Learning spatial-context-aware global visual feature representation for instance image retrieval." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
- [4] Zhang, Qi, et al. "Context-aware attention network for image-text retrieval." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
- [5] Wei, Wenzhang, et al. "Dynamic Visual Semantic Sub-Embeddings and Fast Re-Ranking for Image-Text Retrieval." *IEEE Transactions on Multimedia* (2025).
- [6] Sun, Xiaofei, et al. "Sentence similarity based on contexts." *Transactions of the Association for Computational Linguistics* 10 (2022): 573-588.
- [7] Farouk, Mamdouh. "Measuring sentence similarity: a survey." arXiv preprint arXiv:1910.03940 (2019).
- [8] Ge, Xuri, et al. "Structured multi-modal feature embedding and alignment for image-sentence retrieval." *Proceedings of the 29th ACM international conference on multimedia*. 2021.
- [9] Cui, Zheng, et al. "Context-aware relation enhancement and similarity reasoning for image-text retrieval." *IET Computer Vision* 18.5 (2024): 652-665.
- [10] Wu, Di, Le Zhang, and Yao Chen. "Syntactic-guided optimization of image-text matching for intra-modal modeling." *The Journal of Supercomputing* 81.2 (2025): 367.
- [11] Ji, Zhong, et al. "Hierarchical matching and reasoning for multi-query image retrieval." *Neural Networks* 173 (2024): 106200.
- [12] Baldrati, Alberto, et al. "Zero-shot composed image retrieval with textual inversion." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

- [13] Dong, Jianfeng, et al. "From region to patch: Attribute-aware foreground-background contrastive learning for fine-grained fashion retrieval." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.
- [14] Tang, Yuanmin, et al. "Context-I2W: mapping images to context-dependent words for accurate zero-shot composed image retrieval." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 6. 2024.
- [15] Seonwoo, Yeon, et al. "Ranking-enhanced unsupervised sentence representation learning." arXiv preprint arXiv:2209.04333 (2022).
- [16] Li, Jinhang, and Yingna Li. "A Sentence-Matching Model Based on Multi-Granularity Contextual Key Semantic Interaction." Applied Sciences 14.12 (2024): 5197.
- [17] Gardazi, Nadia Mushtaq, et al. "BERT applications in natural language processing: a review." Artificial Intelligence Review 58.6 (2025): 1-49.
- [18] Li, Bo, Di Liang, and Zixin Zhang. "Comateformer: Combined Attention Transformer for Semantic Sentence Matching." arXiv preprint arXiv:2412.07220 (2024).
- [19] Li, Wei, et al. "CAT-LLM: Context-Aware Training enhanced Large Language Models for multi-modal contextual image retrieval." (2024).
- [20] Lee, Saehyung, et al. "Interactive Text-to-Image Retrieval with Large Language Models: A Plug-and-Play Approach." arXiv preprint arXiv:2406.03411 (2024).