Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-025-00027-1

Original Article

Leveraging Deep Learning for Automated Image Captioning to Improve Search Engine Indexing

Krishna Kumar Mohbey [1], Malika Acharya [1], Shaik Abdul Ahad [1]

1. *Computer Science, Central University of Rajasthan, Ajmer, IND*

Received: May 01, 2025 | Review began: May 26, 2025 | Review ended: March 05, 2026 | Published: March 05, 2026

Abstract

Modern search engines should be able to retrieve images effectively, but traditional techniques mostly rely on manually generated metadata and textual annotations, which are prone to inconsistency and incompleteness. These constraints lower the retrieval accuracy and scalability. This paper presents an image captioning system based on deep learning to produce descriptive, contextually rich captions that facilitate better human image indexing and retrieval. The semantically meaningful captions generated by the proposed system reflect the content and context of a given image. User interaction is provided through a Flask-based web interface, and MongoDB is an extensible, scalable database for storing images and their generated captions. Evaluation of the system was conducted on benchmark image captioning datasets and other image collections to assess its real-world performance. The comparative experiment was conducted against traditional metadata-based retrieval methods using the standard measurement metrics of precision, recall, and mean reciprocal rank. Findings indicate that the caption-based system is superior to traditional techniques in delivering more relevant search results and achieving a higher retrieval accuracy, especially when manual annotations are low or absent. The study shows that deep learning-based image captioning is a viable and scalable alternative to the conventional indexing methods. The system can be used to improve performance in applications such as digital libraries, multimedia archives, and content-intensive web platforms by incorporating semantic understanding directly into the retrieval pipeline, thereby reducing manual input and improving the relevance of retrieved material.

Categories: Search techniques and engines, Image Processing and Analysis, Deep Learning

Keywords: image captioning, image retrieval, deep learning, meacap, search engine

Introduction

The digital environment is evolving rapidly, and the volume of visual content is increasing at a pace search engines have never seen before, posing a significant challenge for them in their quest to index and retrieve images accurately. Historically, traditional methods of image indexing have relied mainly on manually assigned tags, file names, or text, none of which is scalable or comprehensive. Nonetheless, such approaches are not very efficient due to inconsistencies in labelling, the absence of metadata, and scalability issues. In current image search engines, computer vision and natural language processing (NLP) methods are used to address these shortcomings and provide meaningful image representations [1]. Image captioning, which can be defined as the expression of what an image conveys, has received a lot of attention over the last few years. It has a wide range of applications, such as recommendations in editing software, as part of virtual assistants, image indexing, and support to people with disabilities [2]. Nonetheless, image captioning is

Automated captioning offers significant advantages when incorporated into search engine indexing systems. It significantly improves the findability of image searches, increases accessibility for people with visual impairments, supports large-scale content moderation, and reduces the labor-intensive nature of manual tagging. Also, the ability to connect visual and textual information at a high level becomes not only helpful but a necessity as the concept of multimodal search becomes more popular. With the help of neural network structures, namely convolutional neural networks (CNNs) to extract visual features and recurrent neural networks (RNNs) or transformers to generate natural language, these systems can decode intricate visual images and convert them into readings that are easy for humans to understand [6-8]. The latest advancements in deep learning algorithms, including attention models, transformers, and contrastive language-image pretraining, have significantly increased the accuracy of generated captions and the descriptive detail of the content used in multiple online formats. These technologies are expected to change the structure, availability, and use of visual content on many different online platforms [9-10]. Images are processed with CNNs and Vision Transformers [11], whereas textual descriptions are processed with transformer-based NLP models.

The learner of captioning models is agnostic to the intention, and the descriptions generated for images are usually inaccurate and misleading. A conventional image search engine relies on user-created metadata such as file names, alternative text, and user-created tags. However, this has several limitations. To begin with, the lack of descriptive metadata that must be generated manually and applied individually to images in the thousands and millions is time-consuming and demanding. Second, images cannot be considered solitary pictures; instead, they possess embedded semantics. This has been a problem of relying on personal keywords rather than overall contextual interpretation. Images are multidimensional representations that cannot be adequately described by basic tags and that comprise complex relationships among objects, actions, environments, and emotions. For example, a user needs to type a query with certain contextual elements, including: playing (action), child (subject), and park (location), and traditional systems struggle to match this implicit meaning with the correct visual response. Third, traditional systems use conventional keyword-based methods that cannot properly capture the relationships between words and images. This has the disadvantage of requiring users to guess the exact tags or filenames used in the system, which can result in suboptimal search results. This inherent constraint creates a significant gap between human conceptual comprehension and the indexing approaches machines can perform. It is not just a question of inconvenience. In the case of the search query 'sunset over mountains', with traditional systems, the search results are limited solely to images specifically tagged with those phrases, thus ignoring other relevant information that might be labelled as 'dusk in the highlands', 'alpine evening view', etc. This semantic gap will force users to perform multiple searches using different word combinations, severely degrading the search experience and the information discovery process [12].

In this work, we propose an image search engine that uses deep image captioning to improve image search via natural language search. In contrast to classic image search engines that rely on manually set metadata, we will automatically generate captions and use three state-of-the-art models, including Memory-Augmented zero-shot image Captioning framework (MeaCap) [13], Bootstrapping LanguageImage Pre-training (BLIP) [14], and BLIP Large [15]. The key contributions are as follows:

· It enables easy retrieval of images based on natural language queries, rather than using a keyword-based search algorithm to match queries to captions stored in the database, resulting in better retrieval quality.

· It uses a caption-based storage system featuring text indexing to maximize retrieval performance and effectively store millions of images, while also scaling to new models.

· It is highly capable of performing across various fields, such as e-commerce, education, healthcare, and digital asset management, without requiring any domain-specific configurations or metadata.

· The process removes the manual metadata tagging to locate relevant images.

Related works

In this section, we support the premise of image captioning by citing recent works of literature. Mahalakshmi and Sabiyath Fatima [16] combine information retrieval with template generation and text summarization to improve the quality of the summary and image captions. This model summarises written information and produces descriptions of the images of visualised objects. The retrieval system created by Iyer et al. [17] used the VGG16 CNN to generate captions

and Elasticsearch to index images. The goal of the approach is to enhance image retrieval by leveraging semantic information in captions rather than solely on metadata. Vijayaraju [18] incorporates deep learning and image captioning to address the challenges posed by the growing digital world. The system retrieves images with captions most similar to the query, thereby placing greater emphasis on semantic relevance rather than visual similarity alone. Techniques that use encoder-decoder [19], CNN [20], and dynamic grounding words [21,22] are also suggested. An abstract scene graph method framework was proposed by Chen et al. [23] that used the natural semantics and intentions within the graph to produce the captions.

Deep learning has fundamentally changed image captioning by combining computer vision and NLP to generate textual descriptions of pictorial content, as in Chohan et al. [24]. CNNs are typically used for image interpretation, whereas RNNs or Long Short-Term Memory networks are used for language generation. Popular ones include encoder-decoder models and attention mechanisms [25]. The different types of deep learning studied by the researchers include models of visual attention and generative adversarial networks [26].

Many advanced architectures have been proposed to address the problem of image captioning. Ren et al. [27] proposed a generative model that leverages deep recurrent structures to emphasize the importance of jointly combining computer vision and NLP to produce coherent, contextually appropriate sentences. Moreover, the introduction of the Meshed-Memory Transformer framework aims to improve the relationship between picture parts and generated captions, achieving the highest-quality results on benchmark data [28]. The one for all structure provides a unified approach to multimodal tasks and demonstrates that effective image captioning is possible with relatively small datasets.

As noted in recent research, multimodal learning is an important aspect of research in which models are developed to support visual and textual data. An example of this technique is the multimodal RNN, which estimates word-generation probabilities conditioned on visual context [29]. Also, Zhou et al. [30] proposed consolidated Vision-Language Pretraining models that have attracted attention, making them easy to fine-tune across tasks such as image captioning and visual question answering. Such innovations imply a shift toward models of efficiency in visual and textual forms. Hu et al. [31] proposed a large-scale image captioner, LEMON, and conducted the first empirical study of the scaling properties of Vision-Language Pretraining for image captioning. As a result, LEMON establishes new state-of-the-art results across a variety of image captioning benchmarks. In a zero-shot setting, it can generate captions for long-tail visual concepts.

Image captioning is a vital area of computer vision and NLP, where deep learning techniques generate textual captions for visual content. It is important to note that, despite significant advancements, the region continues to face various limitations that prevent the development of robust, effective image captioning systems. One factor that makes image captioning challenging is deep learning models' ability to recognize objects and their relationships in an image. As noted by Hossain et al. [32], many existing methods struggle to understand context, leading to captions that are often incoherent and fail to reflect the image's true nature. The inherent intricacies of language and the contextual relationships between objects pose problems for current models, necessitating further advancements in model structure to enhance understanding and descriptive skills. As noted by Zhang et al. [33], while self-critical sequence training improves performance by aligning training and test metrics, it also introduces issues with reward estimation and dataset dependency, particularly when using popular datasets like MSCOCO. This reliance on datasets underscores the need for diverse training data to improve generalization across different situations.

Materials And Methods

In this paper, we propose an image search engine that leverages deep image captioning to improve image retrieval via natural language queries. Our model automatically produces captions using three state-of-the-art models: MeaCap [13], BLIP [14], and BLIP Large [15]. These models research semantic information in images and produce elaborate written records, which are stored in a MongoDB database for indexing. The search engine part uses Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity to respond to the user's query by searching the captions stored and finding the most similar pictures based on the textual descriptions. Figure *1* illustrates the proposed outline of deep image captioning for search engine indexing (DICSI) architecture.

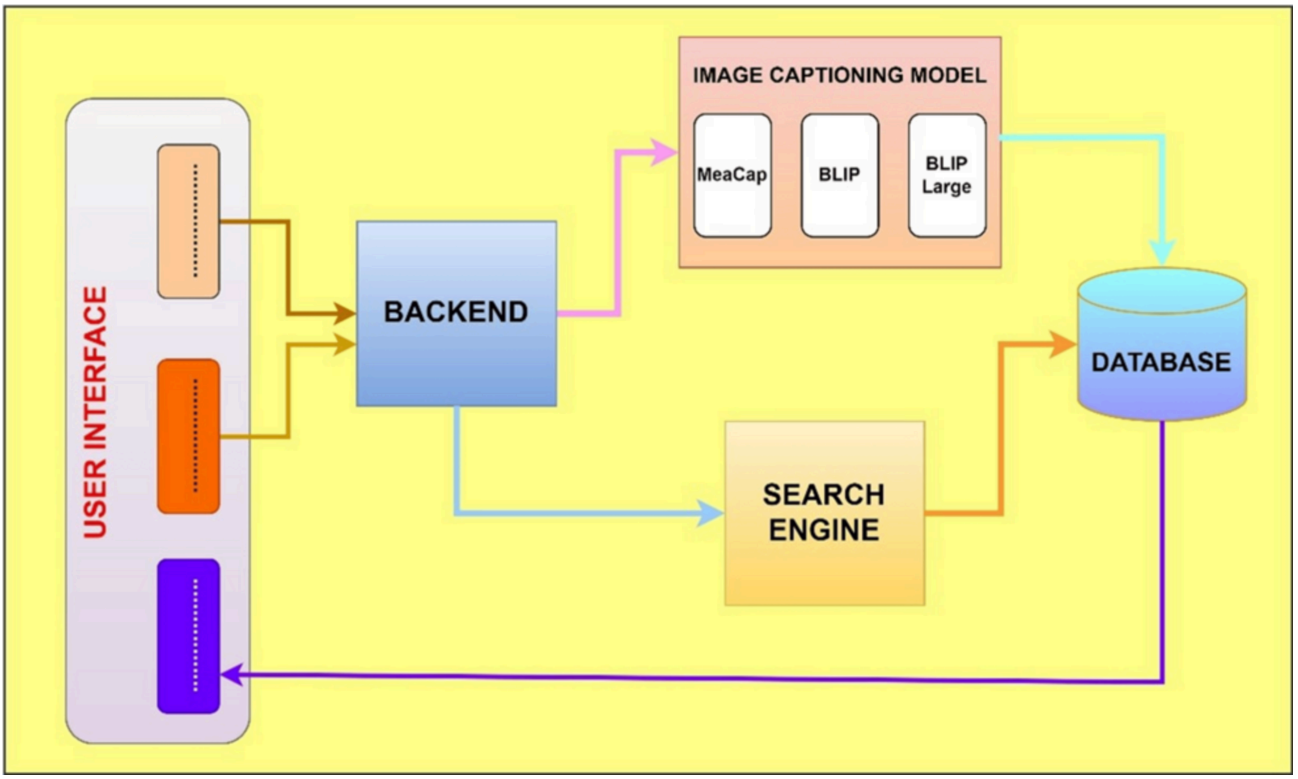

FIGURE 1: DICSI architecture

DICSI, Deep Image Captioning for Search engine Indexing

Proposed model description

Phase 1: Image Captioning

The initial phase of the system uses advanced deep image captioning models to automatically generate textual descriptions for every image in the collection. This stage is essential, as the effectiveness of image retrieval is directly determined by the quality of the captions. The system uses three different captioning models to generate diverse captions. Table *1* compares the three models.

- MeaCap: A transformer-based model that utilizes attention mechanisms to capture detailed image features, producing accurate and contextually aware captions.
- BLIP: A vision-language model that connects visual and linguistic features through language-image pretraining, generating coherent and contextually rich captions for complex images.
- BLIP Large: With its broader design and intensive pretraining, this advanced BLIP version can generate accurate and thorough captions for complex visual scenarios.

Models	Architecture	Strengths	Limitations
MeaCap	Memory-Augmented Transformer with CLIP Guidance	Reduced hallucination, strong semantic alignment	Higher computational complexity in retrieval stage
BLIP	Vision Transformer + Language model	Balanced accuracy & speed	Requires fine-tuning
BLIP Large	Advanced Vision-Language Model	High-quality captions	Computationally expensive

TABLE 1: Comparative analysis of captioning models

The first step in the captioning process is image preprocessing, which includes converting, resizing, and normalizing images for the models. Multiple captions are generated for each image after the three models process it. The model then selects the most descriptive caption based on confidence scores, thereby enhancing accuracy by mitigating individual models' biases and enriching descriptions. Finally, concise yet informative sentences that explain things, events, and context make up the captions.

Phase 2: Caption Storage and Indexing

Upon generating captions, the system archives and organizes them within a MongoDB database. Each image is assigned a distinct ID that corresponds to its caption. The database schema includes:

Image ID: A unique identifier for every image.

Caption: The produced description.

Metadata (Optional): Supplementary details such as image format, resolution, or creation date.

To enhance retrieval efficiency, captions undergo a series of preprocessing steps before indexing:

● Tokenization: Breaking down the caption into individual words or tokens.

● Stopword Removal: Discarding common, insignificant words (e.g., "the," "is," "in") to minimize noise.

● Stemming/Lemmatization: Reducing words to their root forms (e.g., "running" to "run") for improved matching.

● TF-IDF Vectorization: Converting preprocessed captions into TF-IDF vectors, which assign lower weights to frequently occurring words and higher weights to rare, significant words, thereby improving retrieval precision.

After vectorization, captions are indexed alongside their corresponding image references. This inverted index structure allows for rapid lookups, even within extensive datasets. By transforming captions into numerical vectors, the system supports effective similarity evaluations during retrieval. The MongoDB database guarantees scalability, enabling the management of millions of images while ensuring swift query response times.

Phase 3: Image Retrieval Mechanism

This stage emphasizes the image retrieval mechanism, aligning user queries with indexed captions through cosine similarity. This process is crucial for delivering precise and pertinent search outcomes based on natural language descriptions. The retrieval process initiates with query processing. The query is subjected to preprocessing akin to the captions, which involves tokenization, removal of stopwords, and stemming. Subsequently, the preprocessed query is transformed into a TF-IDF vector for comparison against the indexed image captions. Then, we compute the cosine similarity between the query and caption vectors. Images that achieve the highest similarity scores are considered the most relevant. The images are ranked based on the similarity scores, and the relevant top N images are retrieved. The retrieved images are displayed in thumbnail format with captions.

Phase 4: Performance Optimization and Scalability

We have used batch captioning and parallel processing to manage large datasets efficiently. While image batch processing allows the images to be processed in batches and accelerates the process, the parallel processing model saves processing time with multiple GPU thread execution. Further, the indexing features of MongoDB facilitate quick retrieval of captions, even within extensive collections of images. Frequently queried captions are cached to optimize the performance and speed of the retrieval. The web interface for image upload, caption generation, and querying is built using Flask, HTML, CSS, and JavaScript. Algorithm 1 (Figure *2*) demonstrates the image retrieval process using deep learning captioning.

Algorithm 1: Image Retrieval Using Deep Learning Captions

Input: User query (Q), database of images with captions (D)
Output: Ranked list of relevant images R

1. **Preprocess Query**
   a. Tokenize the query Q into individual terms $\{q_1, q_2, ..., q_n\}$
   b. Remove stop words
   c. Apply stemming to each term.
   d. Let Q' = preprocessed query
2. **Vector Representation**
   a. For each term q_i in Q':
      Compute TF-IDF weights
      End For
   b. Represent Q' as a vector V_Q = $\{TF\text{-}IDF(q_1), TF\text{-}IDF(q_2), ..., TF\text{-}IDF(q_n)\}$
3. **Similarity Calculation**
   a. For each image *I* with caption *c* in *D*:
      Compute and store the cosine similarity between (V_Q, V_C) in S(I)
      End For
4. **Ranking and Retrieval**
   a. Sort all images in *D* by their similarity scores S(I) in descending order.
   b. Return the top-k images as the ranked list *R*
5. **Display Results**
   a. Present the ranked images to the user.
   b. display the associated captions and similarity scores

FIGURE 2: Image retrieval using deep learning captions

Deep learning architecture and training configuration

The caption generation component employs three advanced vision-language models: MeaCap, BLIP, and BLIP-Large. MeaCap follows a transformer-based architecture that integrates visual feature extraction with memory-augmented attention mechanisms to enhance semantic alignment and reduce hallucination. BLIP and BLIP-Large are vision-language pretraining models that utilize Vision Transformers as image encoders and transformer-based language decoders to generate coherent and contextually rich captions. To ensure experimental consistency, pretrained model weights were initialized and subsequently fine-tuned on benchmark image captioning datasets. Model training was conducted using the PyTorch framework with the Adam optimizer, an initial learning rate of 5×10^{-5}, and a batch size of 32. Early stopping was applied to prevent overfitting based on validation performance. The dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets. During preprocessing, images were resized to 224 × 224 pixels and

normalized according to each model's input requirements. Data augmentation techniques, including random cropping and horizontal flipping, were applied during training to improve generalization capability. All experiments were performed on a system equipped with an NVIDIA GTX 1650 Max-Q GPU and 16 GB of RAM.

Results

Data source and description

Our image captioning models are trained and assessed using publicly available, open-source datasets to ensure broad applicability and effective generalization. The dataset contains various images across various domains [34]. It is meticulously curated to include visual contexts such as Nature & Wildlife (animals, landscapes), Human Activities (sports, daily life), Urban Environments (buildings, transportation), and Objects & Products (household items, electronics). This diversity enhances the models' ability to generate detailed, contextually relevant captions, thereby improving robustness and effectiveness in image retrieval across diverse domains.

Dataset diversity, generalizability, and bias considerations

The evaluation data consists of images across various categories, such as nature, urban settings, human actions, and consumer products. Such diversity facilitates more generalization, but this does not necessarily reflect culturally specific or specialized areas such as medical imaging. Captioning systems are automated, which means their training data may bias them. For example, models can generalize from demographic characteristics or provide culturally biased accounts. Possible solutions to curb these risks include bias auditing, balanced datasets, and human-in-the-loop checking of sensitive applications in the future.

Experimental setup

The proposed model utilizes Python programming to investigate automated image captioning on a computer equipped with an Intel(R) Core (TM) i7-8550U CPU operating at 1.80GHz, an NVIDIA GTX 1650 Max Q graphics card, and 16 GB of RAM. Deep Learning Libraries such as PyTorch and Hugging Face Transformers are employed for model development. The Flask web framework is utilized to create the user interface and API. MongoDB serves as the database for storing images and their corresponding captions. Additionally, text processing libraries NLTK and Scikit-learn are applied for TF-IDF vectorization and cosine similarity calculation.

Performance evaluation

The performance of the DICSI model has been evaluated against three standard metrics:

Precision@k: The percentage of pertinent images retrieved from the $top - k$ results is measured by precision@k. The process is mathematically represented in Equation (1). It shows the search engine's accuracy in presenting pertinent images in the first k results. A retrieval system with a higher precision@k value is more accurate.

$$\text{Precision@}k = \frac{|\text{Relevant Images in Top-}k|}{k} \qquad (1)$$

Here, k is the number of top-ranked results. It is a user-defined cutoff value (e.g., 1, 5, 10, etc.).

Recall@k: Out of all the relevant images in the dataset, recall@k calculates the percentage of relevant images retrieved (Equation 2). It shows how well the system captures all-important images.

$$\text{Recall@}k = \frac{|\text{Relevant Images in Top-}k|}{|\text{Total Relevant Images}|} \qquad (2)$$

Mean Reciprocal Rank@k: The ranking system's quality is assessed by Mean Reciprocal Rank (MRR), which calculates the speed at which the first pertinent image is retrieved. Across several queries, it determines the average of the reciprocal ranks of the first relevant image. MRR is especially helpful for evaluating how well the system prioritizes pertinent outcomes. Equation 3 represents the MRR@k.

$$\text{MRR} = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|}\frac{1}{\text{rank}_i} \qquad (3)$$

where |*Q*| is the total number of queries, and rank_i is the position of the first relevant image.

Besides the standard metrics for performance evaluation, we have also assessed the performance against Natural Language Generation (NLG) evaluation metrics as discussed below:

Bilingual Evaluation Understudy (BLEU) [35]: BLEU statistics compare n-grams to assess how well the generated caption resembles the reference captions (Equations 4-5). A shortness penalty is applied to deter short captions, and the accuracy of matching n-grams between the generated and reference captions is computed.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp(1-\frac{r}{c}) & \text{if } c \leq r \end{cases} \qquad (4)$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N}\omega_n \cdot \log \rho_n\right) \qquad (5)$$

where *BP* is the brevity penalty, *c* is the length of the generated caption, *r* is the length of the reference caption, ω_n is the weight of n-gram precision, and ρ_n is the precision of the match between n-grams of the generated and reference captions.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [36]: The ROUGE metric calculates the recall of overlapping n-grams between the generated and reference captions. Assessing the completeness of the generated captions works exceptionally well when calculating the percentage of collected reference material. ROUGE-L is the most popular variation for image captioning, which compares the generated and reference captions' Longest Common Subsequence (LCS) as described in Equation 6.

$$\text{ROUGE-L} = \frac{\text{LCS}(\text{gen}, \text{ref})}{\text{len}(\text{ref})} \qquad (6)$$

where $LCS(gen, ref)$ is the length of the longest common subsequence between the generated and reference captions, and $len(ref)$ is the length of the reference caption.

Consensus-based Image Description Evaluation (CIDEr) [37]: Image captioning jobs are the focus of the CIDEr metric. TF-IDF-weighted n-grams are compared to determine the degree of agreement between the generated caption and several reference captions, as represented in Equation 7. CIDEr is more efficient for assessing picture descriptions because it penalizes frequent, meaningless n-grams and concentrates on semantic similarity.

$$\text{CIDEr} = \frac{1}{r}\sum_{i=1}^{r}\frac{\sum_{n=1}^{4}\mathbf{t}_d^{(gen)} \cdot \mathbf{t}_d^{(ref)}}{\|\mathbf{t}_d^{(gen)}\| \cdot \|\mathbf{t}_d^{(ref)}\|} \qquad (7)$$

where *r* refers to the number of reference captions, and $t_d^{(gen)}$ and $t_d^{(ref)}$ represent the TF-IDF vectors for the generated and reference captions.

Caption quality evaluation

Table *2* represents the results of NLG metrics across the three models: MeaCap, BLIP, and BLIP-Large. The BLEU-4 score, which assesses n-gram precision, consistently rises from 0.72 for MeaCap to 0.82 for BLIP Large. This trend indicates that BLIP Large produces captions with word sequences that better match reference captions, reflecting its ability to generate fluent and grammatically correct text while reducing irrelevant or missing words. Similarly, ROUGE-L scores, which evaluate recall-based subsequence matching, follow a similar trajectory. MeaCap scores 0.65, while BLIP Large reaches 0.75, suggesting that these advanced models capture a broader range of reference text. This indicates that BLIP and BLIP Large offer more comprehensive captions, effectively representing the image's essence with greater detail. The higher ROUGE-L score for BLIP Large further highlights its enhanced capability to convey images' contextual and semantic

aspects. The CIDEr scores, which measure semantic similarity using TF-IDF weighting, show the most notable improvement. MeaCap scores 1.10, BLIP achieves 1.25, and BLIP Large leads with 1.35. The consistently high CIDEr scores for both BLIP and BLIP Large demonstrate their effectiveness in generating meaningful and distinctive captions, particularly important for image retrieval tasks, as higher CIDEr scores correlate with more descriptive and content-rich captions, improving search engine performance.

Model	BLEU-4 Score	ROUGE-L Score	CIDEr Score
MeaCap	0.72	0.65	1.10
BLIP	0.78	0.70	1.25
BLIP Large	0.82	0.75	1.35

TABLE 2: Results of caption quality evaluation across MeaCap, BLIP, and BLIP-Large

Search engine performance

The evaluation of the image search engine with 1,000 queries reveals strong retrieval accuracy and effective alignment of natural language queries with pertinent images. A Precision@10 score of 90.5% indicates that, on average, 9 out of the top 10 results are relevant, showcasing the system's ability to prioritize accurate matches and eliminate irrelevant content, making it dependable for precision-sensitive applications such as e-commerce and medical image retrieval. The Recall@10 score of 88.2% demonstrates that the system retrieves nearly 88% of all relevant images within the top 10 results, indicating extensive coverage and reducing the likelihood of overlooking valuable content. This high recall rate is especially advantageous for digital asset management and education, where a diverse array of relevant images is crucial for comprehensive exploration. Furthermore, the MRR score of 0.85 indicates that relevant images generally rank near the top, enabling users to access meaningful content quickly. This underscores the effectiveness of the ranking mechanism, which is vital for efficient and relevant image retrieval in practical scenarios.

Comparative performance with metadata-based retrieval

To further validate the effectiveness of the proposed DICSI framework, we compare its performance against a traditional metadata-based retrieval system using the same dataset and query set. The results are presented in Table *3*.

Method	Precision@10	Recall@10	MRR
Metadata-based	72.4%	68.1%	0.63
DICSI (Proposed)	90.5%	88.2%	0.85

TABLE 3: Comparative performance analysis between traditional metadata-based retrieval and the proposed DICSI framework

DICSI, Deep Image Captioning for Search engine Indexing; MRR, Mean Reciprocal Rank

As shown in Table 3, the proposed DICSI framework significantly outperforms traditional metadata-based retrieval across all evaluation metrics, demonstrating superior semantic alignment and ranking effectiveness.

Qualitative evaluation of caption generation

To balance the quantitative evaluation metrics, a qualitative analysis was conducted to discuss the richness of the semantics and the relevance of the captions' context. The results of MeaCap, BLIP, and BLIP-Large were compared with similar ground-truth annotations and with more traditional metadata based on keywords.

The qualitative evaluation shows that captioning models developed using deep learning generate more context-aware, descriptive representations than the more traditional metadata tags, which are usually single keywords. For example, in a picture of a child running in a park with a dog, the related metadata in the form of the manual consisted of tags such as child, park, and dog. The caption generated by the BLIP-Large model, on the other hand, described the scene as a young child playing with a brown dog in a grassy park in the afternoon. This produced description captures contextual relationships including action (playing), interaction (with a dog), descriptive attributes (brown dog, grass park), and environmental context (time of day), thus providing a more detailed semantic description.

Even though the produced captions were mostly sensible and contextually appropriate given the visual information, certain limitations were occasionally observed. The models tended to provide general descriptions of complex multi-object or culturally sensitive images. These findings indicate that there are still difficulties concerning fine-grained visual reasoning and interpretation of context. Broadly, the qualitative results complement the quantitative ones, which support the fact that the use of captions facilitates semantic matching between user queries and retrieved images.

Discussion

The evaluation of processing time for caption generation and image retrieval shows a trade-off between model complexity and efficiency. The results of the timing analysis are shown in Table 4.

Task	Time Taken (Average)
Caption Generation (MeaCap)	2.5 seconds
Caption Generation (BLIP)	3.0 seconds
Caption Generation (BLIP Large)	4.2 seconds
Image Retrieval	<1 seconds

TABLE 4: Computational performance of different models

Among the three compared captioning models, MeaCap is the fastest, averaging 2.5 seconds per caption; BLIP is next at 3.0 seconds, and BLIP Large at 4.2 seconds. The more time-intensive processing of these models is explained by the greater computational load they require. The BLIP Large is almost 70 times slower than MeaCap, as its architecture is complex and its feature extraction is more advanced. Even though this added time is justified by its high level of accuracy, it creates a scalability problem in real-time applications where speed is a must. The image retrieval process, on the other hand, takes less than 1 second, demonstrating the efficiency of TF-IDF and cosine similarity in matching text-based queries to indexed captions in the shortest possible time. It has a fast response time, so users are almost guaranteed immediate search results, which makes it most applicable in interactive applications like e-commerce and

personal photo management. There is increased efficiency in caption generation with optimization techniques such as batch processing and GPU acceleration. Through parallel computation, such optimizations reduce delays, making the more complex BLIP Large model feasible to use.

Alongside technical performance, the concept of ethical and legal imperatives regarding bias, copyright, and the responsible use of automated captioning systems is raised to facilitate transparency and responsible application.

Integration with real-world search platforms

The presented DICSI framework can be applied to the current search engine architecture using an API-based deployment model. The caption generation module can be implemented as a microservice that provides captions on uploaded photos and stores them in a searchable database. These captions may then be added to both the existing indexing pipelines and the conventional metadata fields. In real-world implementation, the retrieval system may be used as an additional organizational layer, ranking search results based on caption-query similarity scores. In addition, a practical analysis might be conducted using an A/B test, comparing traditional metadata-based retrieval with caption-enhanced indexation to assess changes in click-through rates, user satisfaction, and relevance scores. The architecture allows easy integration into large-scale web platforms and remains compatible with existing infrastructure.

System performance and scalability analysis

To test the system's scalability, retrieval latency was measured across varying dataset sizes-the dataset with 100,000 indexed images returned results in less than a second. Times for caption generation also depended on the model's complexity, with BLIP-Large taking about 4.2 seconds per image. The amount of memory used was proportional to the dataset size, and MongoDB's indexing allowed quick query execution through inverted index structures. These findings indicate that the system performs efficiently in medium-scale applications, with distributed indexing capabilities in large systems.

User interaction and usability evaluation

A controlled user study was conducted to assess practical usability, with 15 participants completing natural language search tasks using the Flask-based interface. Participants were asked to access the appropriate images based on descriptive queries and to compare their experience with traditional keyword search methods. Most participants reported better contextual relevance and correct ranking of results when using a caption-based retrieval system. According to users, generated descriptive captions improved interpretability and reduced the need for query rewording. According to semantic matching, search efficiency was enhanced, especially with multi-attribute or complex queries. Other interface improvements suggested by the participants included better filtering options and result grouping to enhance navigation. Such recommendations will guide subsequent development of the system and larger user-centered assessments.

Conclusions

The study proposes a deep learning-based image captioning system, DICSI, that improves semantic image retrieval and indexing by search engines. The proposed system offers notable improvements over traditional metadata-based indexing and achieves high retrieval effectiveness by automatically generating context-aware captions with MeaCap and BLIP, then adding them to a TF-IDF-based retrieval pipeline. The experimental findings show that the caption-centric strategy achieves Precision@10 of 90.5%, Recall@10 of 88.2%, an MRR of 0.85, and a retrieval latency of less than 1 second. The results indicate that adding semantic understanding to the indexing pipeline improves relevance, especially in low-annotation settings where manual metadata is limited or sporadic. Although this system exhibits good performance, it is limited in some respects. The testing was done using publicly available data and controlled query conditions; a wider real-world implementation could introduce new issues, such as domain shift, significant infrastructure requirements, and bias in caption generation.

Future studies will aim to improve the strength and scalability and to make the proposed DICSI framework more generalizable. A major direction is to combine transformer-based semantic embedding models to replace TF-IDF with deep representation learning to achieve more context-sensitive similarity matching. Furthermore, measuring the system

on domain-specific datasets, such as medical, industrial, or culturally diverse image sets, will be used to assess cross-domain generalization. Future work will also include large-scale user-centered evaluations to measure real-world search satisfaction and usability. Future research will focus on bias-reduction measures, such as balanced datasets and human-in-the-loop validation systems, to make AI use more responsible. Last but not least, the implementation of distributed indexing and cloud-based deployment pipelines will enable real-time, large-scale image retrieval across any enterprise and web-scale application.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Krishna Kumar Mohbey, Malika Acharya, Shaik Abdul Ahad

Acquisition, analysis, or interpretation of data: Krishna Kumar Mohbey, Malika Acharya, Shaik Abdul Ahad

Critical review of the manuscript for important intellectual content: Krishna Kumar Mohbey, Malika Acharya, Shaik Abdul Ahad

Drafting of the manuscript: Malika Acharya, Shaik Abdul Ahad

Supervision: Malika Acharya

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Devi VA, Naved M: Dive in deep learning: computer vision, natural language processing, and signal processing. Machine Learning in Signal Processing. Tanwar S, Nayyar A, Rameshwar R (ed): Chapman and Hall/CRC, New York; 2021. 97-126.
2. Aneja J, Deshpande A, Schwing AG: Convolutional image captioning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. 2018, 5561-5570. 10.1109/CVPR.2018.00583
3. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016, 770-778. 10.1109/CVPR.2016.90
4. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L: Bottom-up and top-down attention for image captioning and visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. 2018, 6077-6086. 10.1109/CVPR.2018.00636
5. Sai Krishna S, Sarath Kumar P, Yaseen M, Rahul Sai P, Raja Sekhar Y, Rout AK: Exploring cutting-edge deep learning advances in image captioning. AIP Conference Proceedings. 2025, 3298:020057. 10.1063/5.0279353
6. Geetha G, Kirthigadevi T, Godwin Ponsam G, Karthik T, Safa M: Image captioning using deep convolutional neural networks (CNNs). Journal of Physics: Conference Series. 2020, 1712:012015. 10.1088/1742-6596/1712/1/012015
7. Singh YP, Ahmed SALE, Singh P, Kumar N, Diwakar M: Image captioning using artificial intelligence. Journal of Physics: Conference Series. 2021, 1854:012048. 10.1088/1742-6596/1854/1/012048
8. Puscasiu A, Fanca A, Gota DI, Valean H: Automated image captioning. 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania. 2020, 1-6. 10.1109/AQTR49680.2020.9129930

9. Wang D, Hu H, Chen D: Transformer with sparse self-attention mechanism for image captioning. Electronics Letters. 2020, 56:764-766. 10.1049/el.2020.0635

10. Lee H, Yoon S, Dernoncourt F, Bui T, Jung K: UMIC: An unreferenced metric for image captioning via contrastive learning. arXiv. 2021, 10.48550/arXiv.2106.14019

11. Elbedwehy S, Medhat T, Hamza T, Alrahmawy MF: Efficient image captioning based on vision transformer models. Computers, Materials & Continua. 2022, 73:1483-1500. 10.32604/cmc.2022.029313

12. Bhoir SV, Patil S: A review on recent advances in content-based image retrieval used in image search engine. Library Philosophy and Practice. 2021, 1-45.

13. Zeng Z, Xie Y, Zhang H, Chen C, Wang Z, Chen B: MeaCap: Memory-augmented zero-shot image captioning. arXiv. 2024, 10.48550/arXiv.2403.03715

14. Li J, Li D, Xiong C, Hoi S: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv. 2022, 10.48550/arXiv.2201.12086

15. Li J, Li D, Savarese S, Hoi S: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv. 2023, 10.48550/arXiv.2301.12597

16. Mahalakshmi P, Sabiyath Fatima N: Summarization of text and image captioning in information retrieval using deep learning techniques. IEEE Access. 2022, 10:18289-18297. 10.1109/ACCESS.2022.3150414

17. Iyer S, Chaturvedi S, Dash T: Image captioning-based image search engine: An alternative to retrieval by metadata. Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing. Bansal J, Das K, Nagar A, Deep K, Ojha A (ed): Springer, Singapore; 2019. 817:181-191. 10.1007/978-981-13-1595-4_14

18. Vijayaraju N: Image Retrieval Using Image Captioning. Master's Projects. 687. 2019. 10.31979/etd.vm9n-39ed

19. Revati Suresh K, Jarapala A, Sudeep PV: Image captioning encoder-decoder models using CNN-RNN architectures: A comparative study. Circuits, Systems, and Signal Processing. 2022, 41:5719-5742. 10.1007/s00034-022-02050-2

20. Iwamura K, Louhi Kasahara JY, Moro A, Yamashita A, Asama H: Image captioning using motion-CNN with object detection. Sensors. 2021, 21:1270. 10.3390/s21041270

21. Jiang W, Zhu M, Fang Y, Shi G, Zhao X, Liu Y: Visual cluster grounding for image captioning. IEEE Transactions on Image Processing. 2022, 31:3920-3934. 10.1109/TIP.2022.3177318

22. Mishra SK, Sinha S, Saha S, Bhattacharyya P: Dynamic convolution-based encoder-decoder framework for image captioning in Hindi. ACM Transactions on Asian and Low-Resource Language Information Processing. 2023, 22:1-18. 10.1145/3573891

23. Chen S, Jin Q, Wang P, Wu Q: Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. arXiv. 2020, 10.48550/arXiv.2003.00387

24. Chohan M, Khan A, Mahar MS, Hassan S, Ghafoor A, Khan M: Image captioning using deep learning: A systematic literature review. International Journal of Advanced Computer Science and Applications (IJACSA). 2020, 11:278-286. 10.14569/IJACSA.2020.0110537

25. Ghandi T, Pourreza H, Mahyar H: Deep learning approaches on image captioning: A review. ACM Computing Surveys. 2023, 56:1-39. 10.1145/3617592

26. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D: Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, 44:3523-3542. 10.1109/TPAMI.2021.3059968

27. Ren S, He K, Girshick R, Sun J: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 39:1137-1149. 10.1109/tpami.2016.2577031

28. Cornia M, Stefanini M, Baraldi L, Cucchiara R: Meshed-memory transformer for image captioning. arXiv. 2020, 10.48550/arXiv.1912.08226

29. Gebru T, Denton R: Beyond fairness in computer vision: a holistic approach to mitigating harms and fostering community-rooted computer vision research. Foundations and Trends® in Computer Graphics and Vision. 2024, 16:215-321.

30. Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J: Unified vision-language pre-training for image captioning and VQA. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34:13041-13049. 10.1609/aaai.v34i07.7005

31. Hu X, Gan Z, Wang J, Yang Z, Liu Z, Lu Y, Wang L: Scaling up vision-language pre-training for image captioning. arXiv. 2022, 10.48550/arXiv.2111.12233

32. Hossain MZ, Sohel F, Shiratuddin MF, Laga H: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR). 2019, 51:1-36. 10.1145/3295748

33. Zhang J, Cao Y, Fang S, Kang Y, Chen CW: Self-critical sequence training for image captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. 2017, 7008-7024. 10.1109/CVPR.2017.742

34. Image dataset. (2024). Accessed: January 5, 2025: https://github.com/yavuzceliker/sample-images.

35. Papineni K, Roukos S, Ward T, Zhu WJ: BLEU: A method for automatic evaluation of machine translation. ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002, 311-318. 10.3115/1073083.1073135

36. Lin CY: Looking for a few good metrics: ROUGE and its evaluation. NTCIR 2004 (NTCIR-4). 2004, 1-8.

37. Vedantam R, Zitnick CL, Parikh D: CIDEr: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. 2015, 4566-4575. 10.1109/CVPR.2015.7299087

How to cite this article:
Mohbey K, Acharya M, Ahad S (March 05, 2026) Leveraging Deep Learning for Automated Image Captioning to Improve Search Engine Indexing . Cureus J Comput Sci 3 : es44389-025-00027-1. DOI https://doi.org/10.7759/s44389-025-00027-1