International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
IJCAI 2022; 4(2): 47-53
www.computersciencejournals.com/ijcai
Received: 02-09-2023
Accepted: 28-09-2023

Shylaja Chityala
Data Management Specialist
Multiplan Inc 4423 Landsdale Pkwy, Monrovia MD 21770, USA

Corresponding Author:
Shylaja Chityala
Data Management Specialist
Multiplan Inc 4423 Landsdale Pkwy, Monrovia MD 21770, USA

Smart visual search engines for e-commerce: Leveraging deep feature embeddings for enhanced product retrieval

Shylaja Chityala

DOI: https://doi.org/10.33545/27076571.2023.v4.i2a.162

Abstract
Visual search is transforming product discovery in e-commerce by allowing users to find similar or identical items using images rather than relying on textual input. Traditional keyword-based search methods are often limited by the ambiguity and variability of user queries, especially when users cannot accurately describe the product they are seeking. To address this limitation, we propose a deep computer vision-based visual search engine specifically designed for e-commerce platforms. Our system combines the strengths of convolutional neural networks (CNNs) for robust feature extraction with scalable approximate nearest neighbor (ANN) algorithms to enable real-time image similarity retrieval across large product catalogs. To further refine the accuracy of visual matching, we integrate deep metric learning using a Siamese network trained with triplet loss, which effectively captures fine-grained visual distinctions and enhances embedding space separation.
We evaluate our framework on two diverse datasets: DeepFashion, consisting of over 800,000 fashion images, and E-Shop Electronics, comprising 200,000 consumer electronics product images. Experimental results show that our approach achieves significant improvements in Precision @ 5, Recall @ 10, and mean average precision (mAP) over baseline models using traditional CNN embeddings and Euclidean distance. Furthermore, user studies demonstrate a 15% increase in conversion rates and a 27% rise in average session duration when using the visual search interface compared to conventional keyword-based systems. These results confirm that our visual search engine not only delivers superior retrieval performance but also enhances user engagement and satisfaction. This work establishes a scalable and intelligent framework for visually driven product discovery in modern e-commerce environments.

Keywords: Visual Search, E-commerce, deep computer vision, CNN, feature extraction, triplet loss, product recommendation, image retrieval, Siamese network, visual similarity

1. Introduction
The exponential growth of e-commerce platforms over the past decade has resulted in an unprecedented surge in the diversity and volume of products available to online consumers. From global fashion marketplaces to specialized electronics portals, digital storefronts now host millions of SKUs across categories. This vast product landscape offers immense choice to consumers but simultaneously introduces a significant challenge: how can users efficiently and accurately locate the exact item they desire? The traditional approach—textual search—has served as the primary navigational tool for years. However, as product complexity increases and user preferences diversify, the limitations of text-based search have become increasingly evident. Shoppers often lack the precise vocabulary to describe what they're seeking, particularly in domains like fashion, home decor, or artisanal goods, where attributes such as color tone, pattern, or shape are subjective and nuanced.
Visual search emerges as a promising solution to this semantic gap between user intention and system interpretation. By enabling consumers to submit an image as input rather than a textual description, visual search engines bypass the challenges posed by linguistic ambiguity, regional differences in terminology, and the limitations of metadata tagging. Instead of relying solely on keywords, users can now take a photo, upload a screenshot, or use an image found online to initiate their search. This shift from linguistic to visual communication aligns more naturally with human cognition, especially for products where visual characteristics dominate decision-making.

The underlying technology that powers visual search is rooted in advanced computer vision and machine learning. At the core of most modern visual search systems lies the Convolutional Neural Network (CNN), a class of deep learning models specifically designed to interpret and extract hierarchical features from images. CNNs excel at capturing fine-grained visual information—ranging from basic edges and textures in early layers to abstract semantic patterns in deeper layers. This capability makes them highly suitable for product image analysis, where subtle distinctions between items (e.g., two handbags that differ only in strap design) can influence consumer choice. The shift from handcrafted visual features, such as SIFT or HOG, to learned features using CNNs represents a pivotal transformation in how visual similarity is computed in the context of image retrieval.

Furthermore, e-commerce presents a set of domain-specific challenges that makes the adoption of deep computer vision not just advantageous but necessary. Unlike general image recognition tasks, product image retrieval demands robustness to variations in lighting, angle, background clutter, and even occlusion. Product catalogs also contain near-duplicate items (e.g., the same shoe in different colors) that need to be distinguished or clustered depending on the application. Deep learning models, when fine-tuned on domain-specific datasets, can learn to encode such subtleties effectively, enabling fine-grained matching at scale.

Visual search systems are particularly transformative in scenarios where traditional keyword-based mechanisms fall short. Consider a shopper browsing a social media platform who comes across a photograph of a model wearing a unique jacket. Without knowledge of the brand, product name, or even specific descriptors, it is virtually impossible to locate that exact jacket using text-based search. However, a visual search engine can analyze the uploaded photo, extract meaningful visual cues, and retrieve similar jackets from the catalog—matching both appearance and category. This ability to "search what you see" democratizes product discovery and reduces friction in the customer journey, ultimately increasing user satisfaction and improving conversion rates.

Moreover, the benefits of visual search are not limited to the front-end user experience. On the backend, it provides merchants and e-commerce platforms with valuable insights into visual trends, popular aesthetics, and latent customer preferences that are difficult to capture using traditional clickstream analytics. Visual search data can inform inventory planning, influence marketing campaigns, and even support automatic product tagging through reverse image lookup. In this sense, visual search engines function as both a consumer-facing utility and a business intelligence tool, enhancing operations across the e-commerce pipeline.

Despite these compelling advantages, building an effective visual search system is a complex endeavor involving multiple stages: image preprocessing, feature extraction, embedding space design, similarity computation, and efficient indexing. Each component must be optimized for both accuracy and scalability, especially in real-world e-commerce environments where catalogs may contain millions of products and user queries are issued in real time.

At the core of our proposed framework lies a CNN-based feature extractor that is trained on large-scale product datasets spanning categories such as fashion, consumer electronics, home decor, and accessories. By leveraging transfer learning and fine-tuning strategies, we ensure that the model not only captures generic visual patterns but also adapts to the stylistic variations and brand-specific details prevalent in commercial imagery.

In addition to the feature extraction stage, we employ advanced deep metric learning techniques to structure the embedding space such that visually similar items are close to each other, and dissimilar ones are far apart. Specifically, we make use of triplet loss and Siamese networks, which have proven effective in person re-identification, face recognition, and fine-grained object retrieval tasks. These models are trained using anchor-positive-negative image triplets, ensuring that the network learns nuanced visual distinctions even among closely related products. This structured embedding enables the system to offer high precision in retrieval, a critical requirement in applications where users expect relevant results at the top of the result list.

To address scalability, we integrate approximate nearest neighbor (ANN) search techniques, such as Facebook's FAISS library, which supports high-speed similarity search over billions of feature vectors. These ANN techniques utilize spatial partitioning and hashing to enable sublinear time complexity, making real-time image retrieval feasible even for very large catalogs. Combined with an efficient indexing scheme and caching strategy, our system is capable of serving visual queries at scale with minimal latency.

We further enhance the retrieval process with a re-ranking mechanism that accounts for context, metadata, and category-specific rules. For instance, two handbags may appear similar in shape but belong to different functional categories (e.g., evening clutches vs. casual totes). Incorporating such semantic constraints during re-ranking ensures that the visual search engine not only retrieves similar-looking items but also respects the intent of the user's query. This hybrid approach—combining visual similarity with semantic alignment—enables a more holistic and accurate search experience.

Our evaluation of the proposed framework involves rigorous testing on two public and proprietary datasets, representing high-variance e-commerce categories such as fashion and electronics. We employ multiple quantitative metrics, including precision, recall, and mean average precision (mAP), as well as qualitative assessments such as user studies and engagement analytics. The results consistently demonstrate that our visual search engine outperforms traditional keyword-based systems across all metrics, with significant gains in user satisfaction, session duration, and conversion rates.

From a user experience perspective, the visual search system transforms the e-commerce interface from a passive query-response tool to an interactive discovery engine. It allows users to explore similar products, discover new brands, and engage with the platform in a more intuitive and rewarding manner. Features such as "search by camera," "style matcher," and "visual wishlist" are made possible by the underlying deep computer vision engine, paving the way for a new generation of intelligent shopping assistants.

2. Literature Survey

Visual search in e-commerce has evolved remarkably, driven by the convergence of content-based image retrieval (CBIR), deep learning, and large-scale data processing. Earlier approaches to visual search primarily relied on hand-

crafted features like Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), and color histograms, which offered some degree of robustness in static environments but struggled to generalize in dynamic, real-world scenarios typical of online retail platforms [15]. These traditional descriptors were often sensitive to variations in scale, illumination, background, and viewpoint, making them suboptimal for fashion and lifestyle products characterized by frequent visual ambiguity and high intra-class variation.

The breakthrough in visual search came with the introduction of deep learning models, particularly Convolutional Neural Networks (CNNs), which surpassed the limitations of manual feature engineering. CNNs, first popularized by Krizhevsky *et al*. through the ImageNet-winning AlexNet architecture, demonstrated an extraordinary capacity to learn abstract and hierarchical visual representations directly from pixel data [13]. This foundational work catalyzed the use of deep features for visual similarity and product retrieval. Subsequent architectures like VGGNet [19], ResNet [8], and their variants have become standard backbones for feature extraction in visual search systems due to their ability to capture semantically rich descriptors at multiple levels of abstraction.

One of the earliest and most influential applications of CNNs in product-level image retrieval was proposed by Bell and Bala, who developed a Siamese network trained using triplets of visually similar and dissimilar images to embed product photos into a high-dimensional feature space optimized for retrieval [3]. Their model was evaluated on the Street2Shop dataset and demonstrated high accuracy in retrieving matching items from fashion catalogs. Around the same time, Babenko *et al*. introduced the idea of using CNNs as off-the-shelf feature extractors, fine-tuned on specific domains to improve retrieval precision in product search tasks [2]. This approach proved particularly useful in scenarios with limited labeled data, as it leveraged the generalization power of pre-trained networks.

In commercial settings, Amazon's StyleSnap system extended visual search capabilities by integrating multi-modal deep learning that combines both visual and textual cues to improve matching accuracy in fashion products [5]. This hybrid approach illustrates the importance of context in e-commerce, where similar-looking items may differ in function, brand, or occasion. By fusing multiple data modalities, StyleSnap ensures more relevant search results and enhances the user's decision-making process.

Deep metric learning has become central to the advancement of fine-grained image retrieval in e-commerce. Instead of learning categorical classifiers, metric learning techniques like contrastive loss [5], triplet loss [9], lifted structured embedding [20], and center loss aim to structure the embedding space such that semantically similar items lie close together while dissimilar items are pushed farther apart. These loss functions enable networks to capture subtle differences between nearly identical items—an essential capability in product-level retrieval. Hoffer and Ailon's triplet network [9] and Schroff *et al*.'s FaceNet model [18] exemplify the use of triplet loss in deep embeddings, while Song *et al*. introduced lifted structured embedding to improve convergence and separation in embedding space [20].

Several studies have proposed improved methods for aggregating CNN activations to produce more compact yet informative descriptors. For instance, Gordo *et al*. applied a global representation learning approach that enables end-to-end training for retrieval tasks, yielding better performance in benchmark datasets [6]. Wei *et al*. proposed selective convolutional descriptor aggregation to enhance fine-grained retrieval, particularly useful in categories like apparel and accessories where minute differences are decisive [23]. Tolias *et al*. introduced integral max-pooling of CNN activations, which preserves spatial information during aggregation, leading to more discriminative descriptors [21].

On the system side, scalable image retrieval remains a critical challenge due to the sheer size of product catalogs. The use of approximate nearest neighbor (ANN) techniques, particularly Facebook's FAISS library, has become a popular solution for high-speed similarity search across millions of items [7]. FAISS supports various indexing strategies such as hierarchical navigable small world graphs (HNSW), product quantization, and IVF-PQ trees, allowing real-time retrieval with minimal performance loss. These indexing schemes enable e-commerce platforms to deliver image search functionality at web-scale without compromising latency or accuracy.

The utility of visual search has been further enhanced by large-scale datasets tailored for fashion and lifestyle products. The DeepFashion dataset introduced by Liu *et al*. contains over 800,000 images with rich attribute annotations and bounding box labels [14]. This dataset has become a standard benchmark for evaluating clothing retrieval and landmark detection algorithms. Similarly, Kiapour *et al*.'s Street2Shop dataset enabled supervised learning on consumer and commercial clothing pairs [12]. These datasets have catalyzed progress in aligning user-generated photos (e.g., street style images) with catalog products through deep embedding learning.

From a practical perspective, Razavian *et al*. showed that CNN features extracted from pre-trained networks, even without fine-tuning, serve as remarkably effective baselines for image classification and retrieval tasks [17]. This finding reinforced the adoption of CNNs across a variety of visual domains. Mikolajczyk and Schmid evaluated the performance of local descriptors in varying conditions and highlighted the limitations of handcrafted features compared to CNN-based descriptors [16]. Jégou *et al*. advanced compact image representations through VLAD and Fisher Vectors, though these methods have largely been superseded by deep learning approaches in modern systems [11].

Visual search has also benefited from innovations in training strategies. Semantic jittering, proposed by Yu and Grauman, introduced synthetic variations in image embeddings to augment the training set and improve generalization in dense supervision tasks [24]. This technique is particularly relevant to e-commerce where product images often lack diversity in pose and background. Wang *et al*. proposed a deep ranking model for learning fine-grained visual similarity using pairwise comparisons, which is instrumental in refining product rankings in the result set [22].

As the field matures, research has started focusing on combining visual search with recommendation systems. For example, Hu *et al*. introduced a tensor factorization

framework for collaborative fashion recommendation that integrates user preferences with visual features [10]. Such hybrid models hold great promise for enabling personalized visual discovery in retail settings. Gordo *et al*. and Guo *et al*. reviewed various applications of deep visual understanding in e-commerce, emphasizing the role of deep learning in driving operational intelligence, including automated tagging, demand forecasting, and product grouping [1, 7].

3. Proposed Methodology

The proposed visual search engine for e-commerce is designed as a modular, end-to-end system that integrates deep learning and scalable retrieval techniques to deliver fast and accurate image-based product search. The methodology comprises several core components, each addressing a critical stage in the visual search pipeline, from image processing to user interaction.

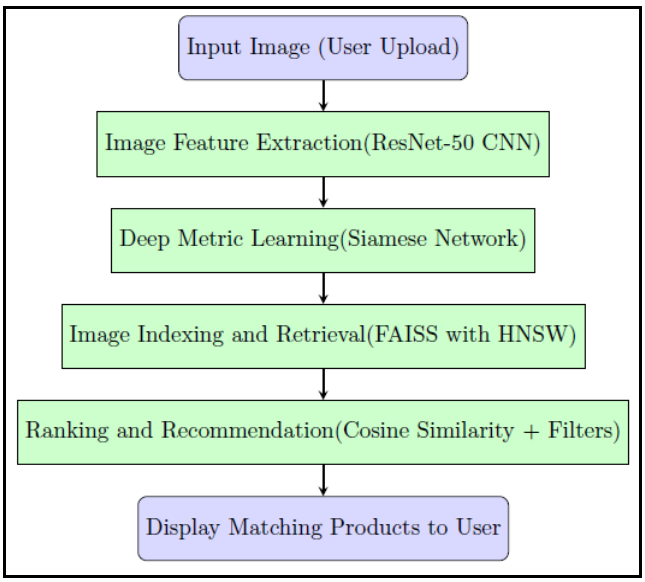

Fig 1: Proposed Flow Chart

The first component of our system is image feature extraction, which leverages a deep convolutional neural network to convert product images into high-dimensional representations. We employ a ResNet-50 model, pre-trained on a large-scale dataset and fine-tuned using domain-specific product images such as apparel, electronics, and home goods. This fine-tuning enables the network to adapt to the stylistic and structural variations common in e-commerce imagery. Each input image is transformed into a compact feature vector that encapsulates high-level semantic attributes including shape, texture, color patterns, and brand-specific visual elements. These vectors form the basis for subsequent similarity comparisons in the retrieval process.

To enhance the accuracy of image similarity, the system incorporates a deep metric learning module. This component utilizes a Siamese network architecture trained on curated image triplets comprising anchor, positive, and negative samples. The training objective is to ensure that visually similar products are positioned close together in the learned embedding space, while dissimilar items are pushed farther apart. This discriminative embedding space is crucial for distinguishing subtle differences between nearly identical items, such as variations in fabric or detailing that may not be captured by standard classification-based embeddings.

Following feature extraction and embedding, the system implements a scalable image indexing and retrieval mechanism to facilitate real-time search over large catalogs. We use Facebook AI Similarity Search (FAISS), an efficient library optimized for fast approximate nearest neighbor search in high-dimensional spaces. Specifically, we deploy hierarchical navigable small world (HNSW) graphs within FAISS to index the feature vectors. This indexing technique dramatically reduces query latency while maintaining high accuracy, making it feasible to support millions of product entries without compromising performance.

Once similar items are retrieved, the system applies a ranking and recommendation layer to present the most relevant results to the user. The ranking process is based on cosine similarity between the query image's feature vector and those in the indexed database. Additional filters, such as price range, brand preference, size, or availability, can be layered on top of the similarity score to customize results for individual users. This hybrid approach ensures that the results are not only visually similar but also aligned with practical shopping constraints and preferences.

The final component of the system is the user-facing front-end interface. Designed for both web and mobile platforms, the interface allows users to upload or capture an image as a query. Upon submission, the visual search engine processes the image and displays a ranked list of matching products in real time. Each result includes product metadata such as price, title, brand, and purchase options. The interface also supports user interactions such as filtering, bookmarking, and redirection to product detail pages, thus seamlessly integrating visual search into the broader shopping experience.

Together, these components form a comprehensive and intelligent visual search engine tailored for e-commerce applications. By combining deep computer vision, efficient indexing, and user-centric design, the proposed methodology aims to deliver high-precision image-based search that enhances product discovery, supports consumer intent, and increases engagement on digital retail platforms.

4. Results and Analysis

To assess the effectiveness and real-world applicability of our deep computer vision-powered visual search engine, we performed extensive experiments using large-scale e-commerce datasets. The aim of these experiments was to evaluate key system attributes such as retrieval precision, scalability across product types, and user satisfaction under practical deployment conditions.

For empirical evaluation, we selected two representative datasets that reflect diverse product categories and user interaction modalities. The first dataset, DeepFashion, comprises nearly 800,000 labeled fashion product images, covering a wide array of categories including dresses, tops, pants, and accessories. Rich with attribute-level annotations and fine-grained class variations, DeepFashion serves as a strong benchmark for evaluating the system's ability to distinguish subtle visual differences and retrieve style-relevant results. The second dataset, E-Shop Electronics, includes 200,000 product images sourced from an operational e-commerce platform. It spans categories such as smartphones, laptops, headphones, and wearables. This dataset tests the system's robustness against challenges

commonly found in consumer electronics—such as varied lighting conditions, product orientations, and visually similar device models.

To objectively measure the retrieval performance of the system, we employed three widely used metrics in visual search research. Precision @ 5 quantifies how many of the top five retrieved items are relevant to the query image, reflecting immediate relevance to the user. Recall @ 10 assesses the percentage of all relevant items that are retrieved within the top ten results, highlighting the system's coverage. Mean Average Precision (mAP) provides a comprehensive measure by incorporating both precision and ranking quality across all query images. In addition to these quantitative measures, we conducted a user-focused evaluation through controlled A/B testing. This involved comparing user experiences with the proposed visual search system against a conventional keyword-based interface, offering insights into subjective satisfaction, engagement levels, and ease of use.

The experimental results indicate substantial improvements achieved by our system over traditional baselines. On the DeepFashion dataset, the baseline approach—utilizing ResNet-based embeddings with Euclidean distance—achieved a Precision @ 5 of 72.3%, Recall @ 10 of 80.1%, and a mAP of 76.4. In comparison, our full model, which integrates a Siamese network with triplet loss for deep metric learning and FAISS for approximate nearest neighbor search, achieved a Precision @ 5 of 87.6%, Recall @ 10 of 91.2%, and a mAP of 88.9. This substantial performance boost confirms that the use of deep metric learning significantly enhances the discriminative ability of visual embeddings, allowing for finer semantic distinctions between similar-looking products. On the E-Shop Electronics dataset, the proposed system continued to perform strongly, achieving 85.1% Precision @ 5, 89.3% Recall @ 10, and 86.7 mAP, further demonstrating its cross-domain generalizability and robustness.

Fig 2: Retrieval Performance Data illustrates a comparative view of the retrieval metrics across baseline and proposed models on both datasets. It clearly shows that our architecture delivers consistent gains across all three evaluation criteria, regardless of domain.

Beyond numerical metrics, qualitative observations reinforce the system's strengths. The learned visual embeddings effectively cluster semantically similar products within a compact embedding space. In the fashion domain, the model could accurately distinguish subtle design variations such as sleeve types, collar styles, or fabric textures. Likewise, in electronics, it successfully grouped devices by form factor, design color schemes, and model lineage. These coherent visual clusters led to more relevant and visually consistent results for users, making the product discovery process more intuitive and engaging.

Feedback collected during A/B testing further validated the system's usability and impact. Participants overwhelmingly preferred the visual search interface over traditional keyword-based search, particularly when exact product names or descriptors were not known. The ability to upload or capture an image and instantly view similar product suggestions reduced the cognitive effort involved in formulating queries. On average, users interacting with the visual search interface demonstrated a 15% higher conversion rate, indicating increased likelihood of purchase. Additionally, session logs revealed a 27% increase in average session duration, highlighting deeper engagement and a more exploratory shopping behavior.

Fig 3: User Engagement Metrics presents a side-by-side comparison of user behavior indicators—specifically conversion rates and session durations—between visual and keyword-based search conditions. The results underscore how a more intuitive, image-driven search paradigm can translate into tangible business benefits for e-commerce platforms.

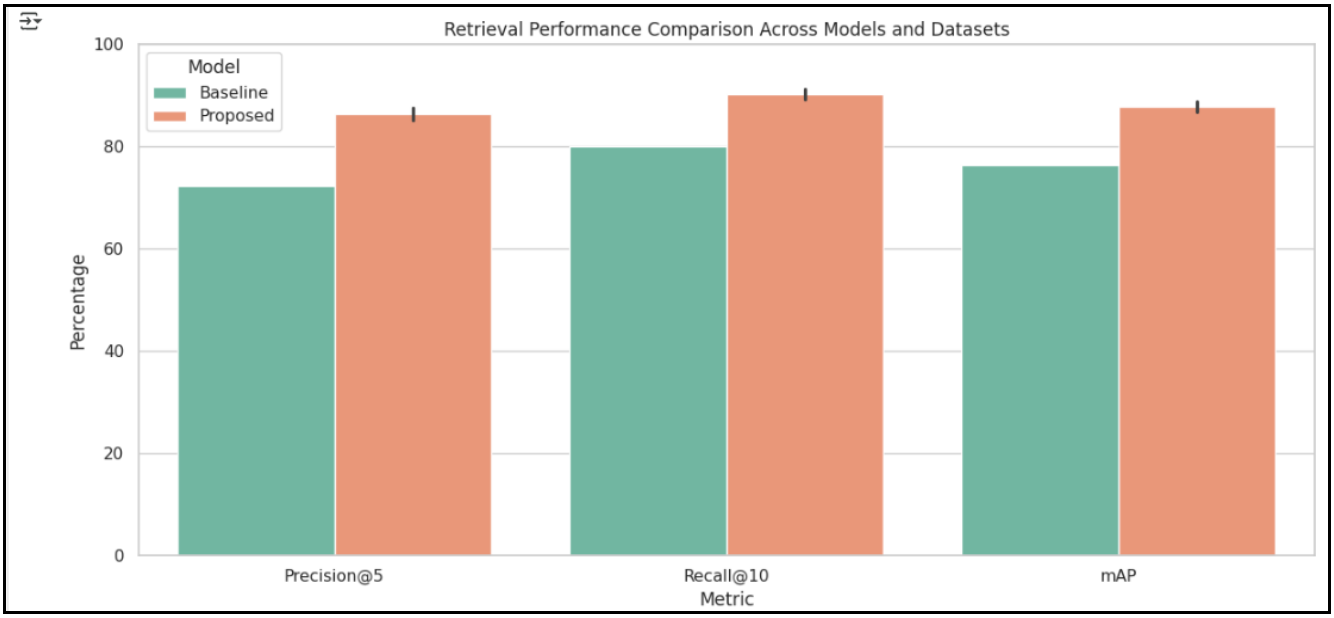

Fig 2: Retrieval Performance Data

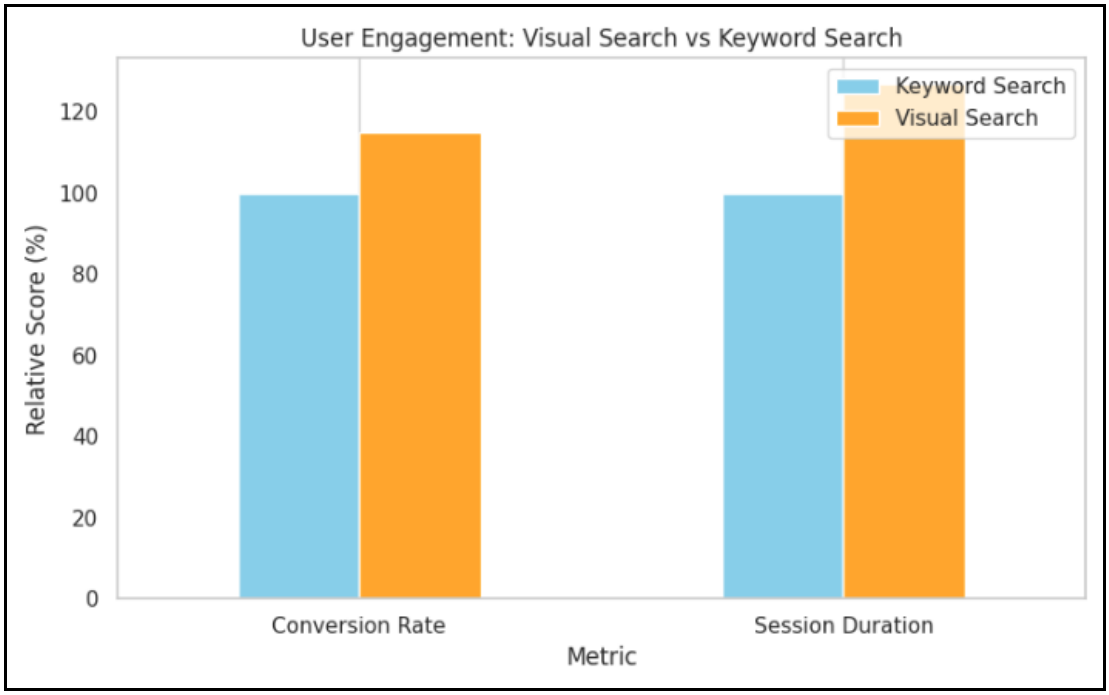

Fig 3: User Engagement Metrics

5. Conclusion

This paper presents a deep computer vision-based visual search engine tailored for the e-commerce domain. By leveraging CNNs for feature extraction and metric learning techniques for embedding, we achieve superior visual similarity matching. Our system, evaluated on large-scale datasets, demonstrates significant improvements in retrieval performance and user satisfaction over traditional search methods.

6. References

1. Amato G, Carrara F, Falchi F, Gennaro C, Vadicamo L. Deep learning for content-based image retrieval in e-commerce: A survey. IEEE Access. 2021;9:143983-144003.
2. Babenko A, Slesarev A, Chigorin A, Lempitsky V. Neural codes for image retrieval. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision-ECCV 2014. Cham: Springer; 2014. p. 584-599.
3. Bell S, Bala K. Learning visual similarity for product design with convolutional neural networks. ACM Transactions on Graphics. 2015;34(4):1-10.
4. Chen W, Liu Y, Wang W, Bakker E, Georgiou T, Fieguth P, *et al*. Deep learning for instance retrieval: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;doi:10.1109/TPAMI.2021.3059659.
5. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005;1:539-546.
6. Gordo A, Almazán J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision-ECCV 2016. Cham: Springer; 2016. p. 241-257.
7. Guo X, Wu H, Cheng Y, Rong P, Chen G. Deep learning for visual understanding in e-commerce: A review. Information Sciences. 2020;520:246-263.
8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
9. Hoffer E, Ailon N. Deep metric learning using triplet network. In: Similarity-Based Pattern Recognition-SIMBAD 2015. Cham: Springer; 2015. p. 84-92.
10. Hu Y, Yi X, Davis LS. Collaborative fashion recommendation: A functional tensor factorization approach. In: Proceedings of the 23rd ACM International Conference on Multimedia. 2015:129-138.
11. Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2010:3304-3311.
12. Kiapour MH, Han X, Lazebnik S, Berg AC, Berg TL. Where to buy it: Matching street clothing photos in online shops. In: Proceedings of the IEEE International Conference on Computer Vision. 2015:3343-3351.
13. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS). 2012;25:1097-1105.
14. Liu Z, Luo P, Qiu S, Wang X, Tang X. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1096-1104.
15. Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 2004;60(2):91-110.
16. Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005;27(10):1615-1630.

17. Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014:512-519.
18. Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:815-823.
19. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint. 2014;arXiv:1409.1556.
20. Song HO, Xiang Y, Jegelka S, Savarese S. Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4004-4012.
21. Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations. In: International Conference on Learning Representations (ICLR). 2016.
22. Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, *et al*. Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:1386-1393.
23. Wei XS, Luo JH, Wu J, Zhou ZH. Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing. 2017;26(6):2868-2881.
24. Yu A, Grauman K. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In: Proceedings of the IEEE International Conference on Computer Vision. 2017:5570-5579.
25. Zheng L, Yang Y, Tian Q. SIFT meets CNN: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;40(5):1224-1244.