10.70711/aitr.v2i9.6879

Research on Multi-Modal Retrieval System of E-Commerce Platform Based on Pre-Training Model

Bingbing Zhang, Yi Han, Xiaofei Han

Xiamen Institute of Technology, Xiamen, Fujian 361024

International Institute of Business Administration, Shanghai International Studies University, Shanghai 200083

Meta Fintech, Menlo Park, CA, USA, 94025

Abstract: In this paper, a multi-modal retrieval system for e-commerce platform is proposed, which integrates three advanced pre-training models: BLIP, CLIP and CLIP Interrogator. The system solves the challenge of traditional keyword-based product search by realizing more accurate and efficient graphic matching. We trained and evaluated our approach using 413, 000 image-text pairs from the Google conceptual Captions dataset. Our method introduces a novel feature fusion mechanism and combines the advantages of several pre-trained models to realize comprehensive visual semantic understanding. The system shows strong performance in daily business scenes and complex artistic product description. Experimental results show that our proposed method can effectively generate detailed and context-aware descriptions and accurately match user queries and product pictures. The adaptability and semantic understanding of the system make it of special value in improving the user experience of e-commerce applications. This research has contributed to the development of intelligent shopping platform by bridging the gap between text query and visual content. It is worth emphasizing that the integration of the CLIP model significantly enhances the e-commerce retrieval system's understanding of user intent and product semantics, thereby making product recommendations more accurate and the search process more targeted.

Keywords: Multi-modal Retrieval; E-commerce; CLIP; BLIP; Image-text Matching

Introduction

With the rapid development of e-commerce, the number of product images on online shopping platforms has grown exponentially. According to Statista, global retail e-commerce sales were estimated to reach about $6.3 trillion in 2023 and are projected to grow nearly $8.5 trillion by 2027, reflecting a compound annual growth rate of approximately 9%. To enhance user shopping experience and search efficiency, developing an effective and accurate image-text retrieval system has become increasingly important. Traditional product retrieval primarily relies on keyword matching, which not only requires merchants to manually add numerous tags but also often fails to accurately capture users' visual needs and products' visual characteristics. In recent years, with the advancement of deep learning technology, particularly the breakthrough in multi-modal pre-trained models, new solutions have emerged for image-text retrieval tasks. Multi-modal pre-trained models can simultaneously understand the semantic information of images and text, establishing connections between the two modalities to achieve more precise image-text matching. This technology has broad application prospects in the e-commerce field, not only improving user shopping experience but also helping merchants better showcase products and increase sales efficiency. In e-commerce scenarios, users often describe products they want to purchase using natural language, such as "a beige knit cardigan" or "a minimalist style desk." Traditional keyword matching methods might miss many relevant products, while retrieval systems based on multi-modal pre-trained models can understand the semantic content of text descriptions and find the most matching product images in the visual space. This research aims to build such an intelligent image-text retrieval system to enhance the search experience on e-commerce platforms.

1. Literature Review

In recent years, multi-modal pre-trained models have made significant progress in image-text retrieval. Radford et al. [1] proposed the CLIP model, which pioneered the approach of learning visual models through natural language supervision. Through large-scale image-text pair training, they achieved excellent zero-shot transfer capabilities. Subsequently, Li et al. [2] introduced the BLIP model, which achieved breakthrough progress under a unified vision-language understanding and generation framework through an innovative bootstrapping language-image pre-training strategy. In the field of multi-modal retrieval applications for e-commerce, Gu et al. [3] were among the early ex-

plorers of multi-modal and multi-domain embedding learning in fashion product retrieval, proposing an embedding learning framework that comprehensively considers visual, textual, and attribute features. Jin et al. [4] addressed the unique characteristics of e-commerce scenarios by proposing an instance-level multi-modal pre-training method for large-scale applications, significantly improving retrieval performance in e-commerce settings. Regarding systematic research on cross-modal retrieval methods, Wang et al. [5] conducted a comprehensive review of existing cross-modal retrieval methods, outlining the field's research progress and indicating future directions. Yu et al. [6] proposed the Heterogeneous Attention Network, effectively addressing the modal disparity issues in cross-modal retrieval and improving both retrieval efficiency and accuracy. In terms of model optimization and practical deployment, Ji et al. [7] introduced an online distillation-enhanced multi-modal Transformer model, optimizing model performance through sequential recommendation approaches to better suit practical application scenarios. This research not only improved model efficiency but also provided important references for deploying multi-modal models in real-world systems.

2. Data and Model Introduction

2.1 Data Introduction

This study uses a large-scale image-text dataset from Kaggle, sourced from Google's Conceptual Captions, containing around 413, 000 image-description pairs. The dataset features diverse real-world content aligned with e-commerce domains, and all text descriptions are manually cleaned to ensure quality. Images range from 224×224 to 512×512 pixels, suitable for training. During preprocessing, images were resized and normalized, while text data underwent tokenization and stop word removal. These steps enhanced model training efficiency and ensured consistency across multimodal inputs for the vision-language retrieval system.

2.2 Model Architecture

This research integrates three advanced pre-trained models—BLIP, CLIP, and CLIP Interrogator—as core components to enhance retrieval performance. BLIP employs a ViT-based image encoder and BERT-based text encoder to align visual and textual features via contrastive learning. It excels in capturing long-range dependencies and deep semantic relationships. CLIP, developed by OpenAI and trained on 400 million image-text pairs, uses a dual-encoder architecture and offers strong zero-shot capabilities. In our system, CLIP extracts high-level semantic features, supporting accurate image-text matching. This multi-model fusion leverages the strengths of each model to ensure robust cross-modal understanding.

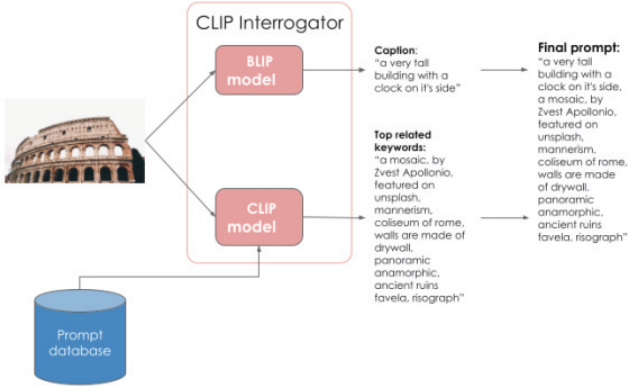

Figure 1. CLIP Interrogator Architecture and Workflow

As shown in Figure 1, the CLIP Interrogator combines both BLIP and CLIP models to generate comprehensive image descriptions by extracting captions and related keywords from the input image, which are then merged into a final detailed prompt. CLIP Interrogator is an important extension based on the CLIP model, specifically designed for generating detailed image descriptions. This model can extract rich visual features from images, including information about objects, scenes, and styles across multiple dimensions. In our system, CLIP Interrogator serves as a bridge, converting visual features from images into text descriptions, thus providing additional semantic information for image-text matching.

For an input image I and text query T, the feature extraction process can be represented as:

$$F_B^I = BLIP_{img}(I) \in R^{d_B} \qquad F_B^T = BLIP_{txt}(T) \in R^{d_B}$$

$$F_C^I = CLIP_{img}(I) \in R^{d_C} \qquad F_C^T = CLIP_{txt}(T) \in R^{d_C}$$

Where F_B^I, F_B^T are BLIP image and text features, and F_C^I, F_C^T are CLIP features, respectively. d_B and d_C represent their respective feature dimensions. The CLIP Interrogator generates text description D from image I:

$$D = Interrogator(I) = \{w_1, w_2, \ldots, w_n\}$$

The final similarity score between image and text is computed using:

$$sim(I,T) = \frac{F^I_{fused} \cdot F^T_{fused}}{|F^I_{fused}| \cdot |F^T_{fused}|}$$

The collaborative workflow of these three models is as follows: First, when a user inputs a text query, the system uses both BLIP and CLIP to extract semantic features from the text. Meanwhile, for product images in the database, the system uses CLIP Interrogator to generate detailed descriptions and combines these with image features extracted by BLIP and CLIP. Finally, the system calculates similarity between feature vectors to return the most matching product images. This multi-model fusion approach effectively utilizes the advantages of each model to provide more accurate retrieval results.

3. Model results analysis

To validate the effectiveness of our proposed multi-modal retrieval system, we analyzed two representative test cases. As shown in Figure 2, the system demonstrates strong capabilities in image understanding and description generation across different scenarios. Figure 2 showcases the system's ability to comprehend everyday commercial scenarios. Given the original input describing a scene at a donut shop, the system-generated description accurately captures key elements of the scene, including "a man standing in front of a counter," "digital rendering," and "donut." This demonstrates our system's effectiveness in understanding visual elements and spatial relationships in commercial settings.

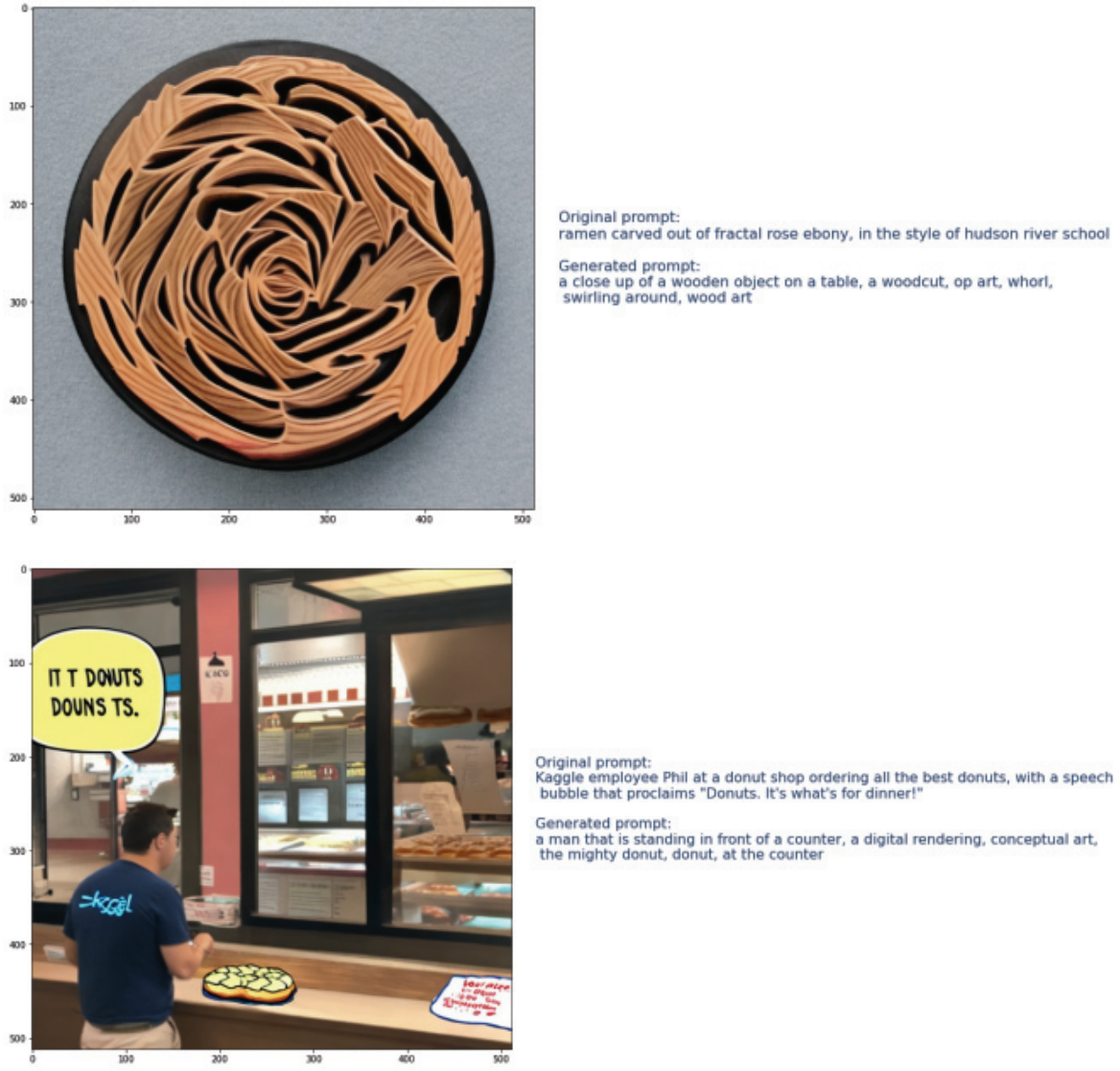

Figure 2. Text Generation Example in Donut Shop Scene and Detailed Art Description Generation Example

Figure 2 highlights the system's ability to process complex artistic content, such as a wooden rose carving, by identifying both physical attributes and artistic features like "woodcut" and "op art." These results demonstrate the system's strong adaptability across scenes, from everyday items to artistic products. It effectively captures objects, spatial relations, and visual styles, showing deep semantic understanding. The system's capacity to generate context-aware, multi-dimensional descriptions enables accurate image-text matching, making it highly valuable for e-commerce applications where precise product retrieval and rich content interpretation are essential for enhancing the online shopping experience.

4. Conclusion

In this paper, we have presented a multi-modal retrieval system for e-commerce platforms that leverages advanced pre-trained models. Through the innovative integration of BLIP, CLIP, and CLIP Interrogator, our research has successfully developed a comprehensive solution that effectively bridges the gap between textual queries and visual content in online shopping scenarios This system helps users to retrieve matching accuracy during the shopping process, improve the shopping experience, and reduce search abandonment, retention, etc. Our research makes several significant contributions to the field of multi-modal retrieval. We have proposed a novel multi-model fusion approach that effectively combines the strengths of three state-of-the-art pre-trained models, leveraging BLIP's powerful vision-language understanding, CLIP's zero-shot learning capabilities, and CLIP Interrogator's detailed description generation abilities. The feature fusion mechanism we developed achieves more comprehensive and accurate representations of both images and text, enabling better cross-modal understanding and matching through the weighted combination of features from different models. The experimental results have validated the effectiveness of our approach, demonstrating that the system can successfully generate accurate and detailed descriptions of product images, understand complex visual features, and provide reliable image-text matching for e-commerce applications. Our system shows strong adaptability across diverse e-commerce scenarios, from everyday products to artistic items, generating detailed, context-aware descriptions that facilitate accurate product matching. Looking forward, several promising directions for future research emerge from this work. While our current system demonstrates strong performance, there is potential for further efficiency optimization to reduce computational complexity while maintaining performance, making the system more suitable for large-scale deployment. Future work will focus on model compression, domain-specific fine-tuning, and real-time performance. Integrating our system with recommendation engines and AR/VR technologies could further enhance immersive shopping experiences. This research lays a strong foundation for scalable, accurate, and intelligent multimodal retrieval in the future of e-commerce.

References

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[2] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International conference on machine learning. PMLR, 2022: 12888-12900.

[3] Gu X, Wong Y, Shou L, et al. Multi-modal and multi-domain embedding learning for fashion retrieval and analysis[J]. IEEE Transactions on Multimedia, 2018, 21(6): 1524-1537.

[4] Jin Y, Li Y, Yuan Z, et al. Learning instance-level representation for large-scale multi-modal pretraining in e-commerce[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11060-11069.

[5] Wang T, Li F, Zhu L, et al. Cross-modal retrieval: a systematic review of methods and future directions[J]. Proceedings of the IEEE, 2025.

[6] Yu T, Yang Y, Li Y, et al. Heterogeneous attention network for effective and efficient cross-modal retrieval[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021: 1146-1156.

[7] Ji W, Liu X, Zhang A, et al. Online distillation-enhanced multi-modal transformer for sequential recommendation[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 955-965.