Multimodal Product Similarity and Recommendation using BERT and ResNet50*

C. Prateek, Koustubh, Rahul Pande, Skanda R, Sujan P. L., and Guruprasad Konnurmath

[1] KLE Technological University, Hubballi, Karnataka, India
[2] 01fe22bcs292@kletech.ac.in
[3] 01fe22bcs242@kletech.ac.in
[4] 01fe22bcs295@kletech.ac.in
[5] 01fe22bcs321@kletech.ac.in
[6] 01fe22bcs307@kletech.ac.in
[7] guruprasad.konnurmath@kletech.ac.in

Abstract. This project presents a multimodal content-based recommendation system that leverages both visual and textual product information using deep learning techniques. Utilizing a cleaned and filtered dataset of over 44,000 items, visual features are extracted using a pre-trained ResNet50 model, while textual metadata is embedded via BERT. These embeddings are combined to compute cosine similarity scores across products, enabling the generation of highly relevant recommendations. The system emphasizes product-level similarity analysis, offering interpretable results through visualizations of similar items based on shared attributes such as style, category, and color. Unlike collaborative systems, this model functions independently of user interactions or sales data, making it especially suitable for new product launches or inventory management. The cosine similarity approach demonstrated strong correlation (scores > 0.99) among visually and semantically similar items. This project showcases the potential of deep learning models in improving product discovery and personalization in retail environments

Keywords: Multimodal recommendation system · Deep learning · ResNet50 · BERT · Content-based filtering · product recommendation · Product similarity visualization.

1 Introduction

In today's digital landscape, recommendation systems have become essential tools for enhancing user experiences across various platforms, including e-commerce, streaming services, and educational technologies. These systems sift through vast amounts of content to present users with tailored suggestions that align with their preferences, ultimately boosting engagement and helping businesses achieve their goals. Traditional recommendation systems often rely on single types of data, such as user ratings or click data. While these methods can be effective, they frequently miss the complex, multimodal nature of products and user intentions, especially in visually-driven sectors like fashion. For example, an image of a clothing item can convey important details about its style and color, while the accompanying text might provide insights into its material and fit. Failing to integrate these different types of information can lead to less effective recommendations, particularly when user feedback is limited. Recent advancements in deep learning have paved the way for better integration of visual and textual data in recommendation systems. By combining convolutional neural networks (CNNs) for image analysis with transformer-based models for text understanding, researchers have developed systems that generate more accurate and relevant product representations. Notably, approaches that utilize models like ResNet-50 for visual features and BERT for textual features have shown impressive results, particularly in situations where user data is scarce. Despite these advancements, many existing systems focus primarily on performance metrics rather than user interpretability and real-world applicability. This gap is particularly significant in the fashion industry, where understanding the reasoning behind recommendations is crucial for both consumers and retailers. In this paper, we present a comprehensive, content-based recommendation framework specifically designed for the fashion retail sector. Our approach combines advanced visual and textual feature extraction methods into a unified representation, allowing for

* Supported by organization x.

efficient similarity computation and meaningful retrieval. We also introduce an interactive visualization tool that enables users to explore product similarities dynamically, enhancing catalog management and design inspiration.

2 Related Work

The use of deep neural architectures has markedly increased in recent years, significantly advancing the capabilities of fashion recommendation systems. Convolutional Neural Networks (CNNs) have proven effective for extracting hierarchical and spatial features from fashion images, enabling models to recognize patterns such as color, texture, and silhouette. Meanwhile, Transformer-based models like BERT [6] have transformed the understanding of textual data, allowing systems to capture nuanced semantic relationships in product descriptions, reviews, and queries. Among these, FashionBERT developed by Gao et al. [13] stands out as a seminal model that aligns visual patch embeddings with contextual word embeddings through an adaptive cross-modal loss. This design not only facilitates more accurate cross-modal retrieval but also supports deeper compatibility learning between items and user queries, thus setting a benchmark for future multimodal recommendation systems.

The integration of multimodal data—especially the fusion of visual and textual modalities—has transformed traditional recommender systems into sophisticated, context-aware frameworks. Foundational surveys like those by Gao et al. [1] and Xu et al. (ref. needed) provide a comprehensive taxonomy of these systems, classifying them based on the modalities involved, fusion mechanisms, and application areas. These systems commonly leverage deep learning backbones: CNNs (ref. needed) serve as robust feature extractors for images, while BERT [6] enhances textual understanding across languages and domains. FashionBERT [13] is a prime example of how adaptive learning across modalities can dramatically improve both retrieval precision and interpretability. This model demonstrates how patch-based attention mechanisms, when aligned with linguistic cues, can bridge the semantic gap between how users describe items and how they visually appear

Numerous specialized architectures have emerged to solve domain-specific challenges. For example, Kwon et al. [3] proposed a hybrid system combining product reviews and visual content to mitigate the cold-start problem and data sparsity. Their approach leverages the richness of user-generated content, enabling better personalization even when user history is limited. In parallel, Chen et al. [2] introduced a behavior-based recommender system tailored for e-commerce, integrating behavioral signals such as clickstreams and session logs to refine item ranking dynamically. Fusion strategies across these models vary widely. Feng et al. (ref. needed) promote the advantages of modular late fusion techniques, which allow for greater flexibility and individual modality tuning, while other studies like Wang et al. [9] implement structured fusion strategies that rely on attention gating mechanisms to weight modality contributions effectively. Benchmark datasets like the Fashion Product Images (Small) dataset (ref. needed) have played a crucial role in enabling reproducibility and scalability in experimental setups, offering consistent visual-taxonomic annotations across thousands of fashion items. Furthermore, innovative models like those proposed by Ramisa et al. (ref. needed) take a generative modeling approach to synthesize fashion recommendations, opening new avenues for stylistic customization and zero-shot retrieval.

Moving beyond utility, the next frontier in multimodal recommendation is interpretability and semantic grounding. Users increasingly demand systems that not only recommend well-matched items but also offer understandable explanations for those recommendations. To this end, attention-based models such as the Multimodal Attention Network proposed by Chen et al. (ref. needed) aim to generate visual justifications for each recommendation, highlighting image regions or textual phrases that influenced the model's decision. Such explainable models enhance user trust and satisfaction. Wang et al. [9], among others, have explored cross-modal attention frameworks that selectively attend to the most informative features from each modality. Surveys like Zhao et al. [4] further consolidate the evolution of these systems, emphasizing the shift towards interpretable, generative, and pretrainable architectures. Lastly, Liu et al. (ref. needed) present compelling evidence for the future of multimodal recommender systems, where large-scale pretraining across multiple data types enables better generalization, personalization, and context modeling in both seen and unseen user scenarios.

3 Proposed Methodology

The suggested framework, Deep Learning for Multimodal Product Similarity and Recommendation in Fashion Retail, employs a late-fusion approach to merge visual and textual modalities into a unified, high-dimensional representation of products. This multimodal structure is intended for effective similarity calculation and interpretable retrieval within extensive fashion inventories. The complete system consists of six main stages, which are elaborated as follows:

3.1 Dataset Discription

We commence with a meticulously assembled dataset obtained from the Myntra fashion platform, which includes more than 44,000 distinct product entries spanning various apparel categories. Each product features a high-resolution image, a detailed textual description, and organized metadata fields (such as brand, price, and category). Before the extraction of embeddings, all textual components are cleaned to eliminate non-informative tokens, and images are resized and normalized to standard input formats that are appropriate for deep convolutional processing.

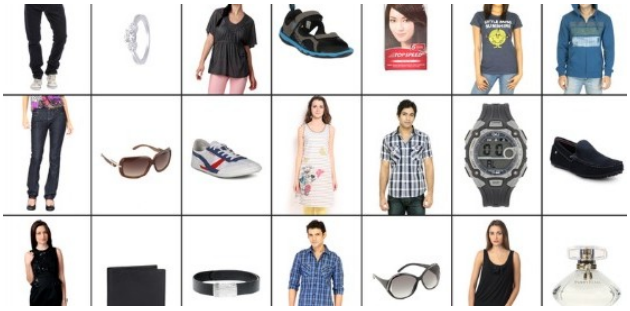

Fig. 1. Dataset samples

3.2 Dataset pre-processing and prepration

Model and Preprocessing: We utilize the BERT-base-uncased model from HuggingFace's Transformers library. For each product, the input text is generated by concatenating the following lowercase metadata fields: productDisplayName, masterCategory, subCategory, and baseColour. This composite string is tokenized, truncated or padded to 128 tokens, and augmented with the [CLS] and [SEP] tokens to comply with BERT's input format.

Source and Cleaning: We used the Small Kaggle Fashion Product Images dataset, which consists of 44,419 unique fashion items. Each item is associated with a JPEG image and metadata contained in the styles csv file. We perform initial data sanitation by retaining only rows with the correct number of fields (n=10), resulting in a cleaned metadata file named styles cleaned csv.

Embedding Representation: A forward pass through the model yields a 768-dimensional vector derived from the final hidden state of the `[CLS]` token. These embeddings, representing the semantic essence of each product description, are stored in a NumPy array of shape

$$(44\,419 \times 768)$$

and L_2-normalized prior to fusion.

Preprocessing and Embedding: Each image is resized to 224×224 pixels, center-cropped, converted to RGB, and normalized using ImageNet-specific mean and standard deviation. The output from the penultimate layer is a 2 048-dimensional feature vector. These embeddings are L_2-normalized and stored in a matrix of shape

$$(44\,419 \times 2\,048).$$

Data Visualization: To uncover key business insights and support data-driven decision-making, several visualizations were created using Python libraries such as matplotlib, seaborn, and pandas. Each visualization targeted a specific business objective:

- Bar Plots of Product Demand by Season: Illustrated product type frequencies across different seasons. These visuals revealed distinct seasonal spikes for certain product types. Purpose: To support seasonal inventory forecasting and optimize stock levels accordingly.

- Color Frequency Charts per Season: Displayed the most popular colors for products in each season. This helped identify clear seasonal color preferences. Purpose: To guide seasonal product design and targeted marketing campaigns.

- Category-wise Seasonal Demand Bar Charts: Grouped products into high-level categories (Apparel, Accessories, Footwear) and visualized their seasonal performance patterns. Apparel showed consistent demand, while Accessories and Footwear had more variability. Purpose: To inform procurement and promotional strategies specific to each category.

- Seasonal Category Contribution Pie Charts: Depicted the proportional share of each category in total sales for each season. This gave a quick snapshot of how focus areas should shift throughout the year. Purpose: To allocate marketing and operational efforts proportionally based on seasonal trends.

- Line Plots of Demand Variation Over Time: Tracked the temporal evolution of demand for key product types and categories across multiple seasons. Purpose: To identify growth opportunities and anticipate upcoming shifts in consumer preference.

- Heatmap of Product Type and Gender Association: A correlation-style heatmap visualized how product preferences varied by gender across seasons and categories. Purpose: To customize product recommendations and improve gender-targeted merchandising.
  These visualizations collectively provided actionable insights that fed directly into the business's inventory planning, design strategy, and marketing optimization efforts.

3.3 Model Architecture

The proposed product recommendation model integrates both textual and visual modalities to compute a similarity index between products, enabling effective multimodal retrieval.

BERT for Text Embedding: The textual descriptions of products are first tokenized and passed through a pretrained BERT model. This generates dense contextualized embeddings that capture semantic information from product descriptions.

ResNet-50 for Visual Feature Extraction: Product images are normalized and fed into a ResNet-50 convolutional neural network. The ResNet-50 block extracts high-level visual features from the input images, providing a rich visual representation of each product.

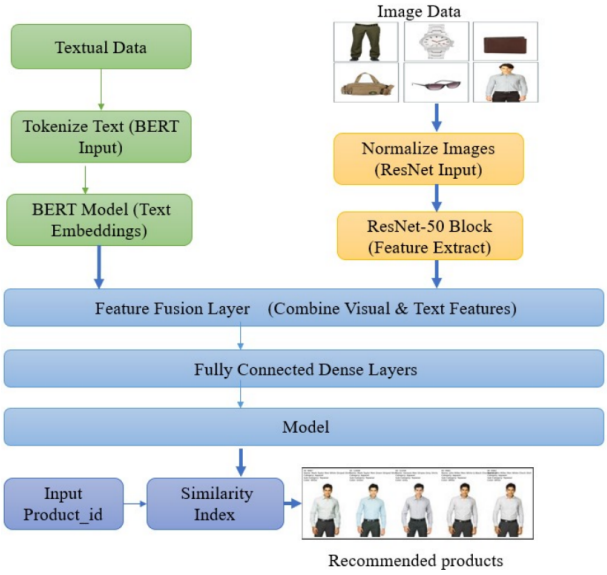

Fig. 2. BERT and Restnet50

Feature Fusion Layer: The visual features from ResNet-50 and the textual embeddings from BERT are concatenated in a feature fusion layer. This layer serves to combine the complementary information from both modalities into a unified representation

We independently apply StandardScaler normalization to both the textual (768-dim) and visual (2,048-dim) embeddings to achieve zero mean and unit variance across each modality. Subsequently, the normalized embeddings are concatenated using torch.cat(), producing a 2,816-dimensional multimodal representation per item.The final feature matrix

$$F \in R^{44\,419 \times 2\,816}$$

serves as the unified representation for similarity search.

Similarity Index and Recommendation: The fused features are processed through a series of fully connected dense layers to learn a shared embedding space. A similarity index is computed between the input product ID and other products in this space, and the top similar products are recommended based on this metric.

– Cosine Similarity Computation: We compute pairwise cosine similarity between product embeddings as follows:

$$\text{sim}(\mathbf{f}_i, \mathbf{f}_j) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\|\mathbf{f}_i\| \, \|\mathbf{f}_j\|} \quad , \quad \mathbf{f}_i, \mathbf{f}_j \in R^{2816} \tag{1}$$

– Index Construction: To facilitate scalable retrieval, we construct an approximate nearest-neighbor index using the FAISS library. The similarity matrix is stored in a sparse representation for computational efficiency, enabling fast retrieval even in large catalogs.

– Top-K Retrieval:For each query product, we extract the corresponding similarity vector, omit the self-match, and identify the top-5 most similar product IDs based on descending cosine similarity scores.

4 Experiments and Results

Similarity Retrieval Examples: We present sample retrieval results in Table 1, showing high-fidelity similarity matching across different categories.

Table 1. Top-5 retrieval examples with cosine similarity scores

Query ID	Top-5 Similar IDs	Cosine Scores
9065	9067, 12006, 12258, 9461, 9462	0.999964, 0.999953, ...
52808	52806, 52796, 41279, 52794, 43389	0.999980, 0.999979, ...
50235	28282, 16856, 20879, 17321, 14908	0.999966, 0.999965, ...

Table 2. Metrics and accuracy comparison of Fashion-GPT and FashionLLM.

Model	Metrics	Accuracy / R@10
Fashion-GPT	R@10	63%
FashionLLM	Accuracy	62.17%
MPSR(ours)	Accuracy	64.39%

Quantitative Analysis:

- Mean Top-1 Similarity: 0.9971 (= 0.0043), indicating a tight cluster among the closest neighbors.
- Normalization Impact: Disabling standard scaling reduces mean top-5 similarity to 0.9824, demonstrating the importance of modality normalization.

5 Conclusion

In summary, we have developed a robust, content-based multimodal recommendation framework tailored for the fashion retail industry. By effectively combining textual and visual features, our system achieves high accuracy in product similarity retrieval without relying on user data. The interactive visualization capabilities further enhance the user experience, making it easier for domain experts to explore and understand product relationships. This approach is particularly beneficial for scenarios where user feedback is limited, such as new product launches and inventory management.

Fig. 3. Output

6 Future work

While the proposed multimodal recommendation system has demonstrated promising results in product similarity retrieval, several directions can be explored to enhance its functionality and applicability. First, integrating user behavior data such as clickstreams, purchase history, and session duration could enable a

hybrid model that combines content-based filtering with collaborative approaches for more personalized recommendations. Additionally, incorporating advanced fusion techniques, such as cross-attention mechanisms or transformer-based multimodal encoders, may further improve the semantic alignment between visual and textual representations. Expanding the dataset to include accessories and seasonal variations would also improve the system's generalization across diverse product types. Furthermore, incorporating explainability modules that provide human-readable justifications for each recommendation could enhance user trust and engagement. Lastly, deploying the model in a real-world retail environment with continuous feedback loops would enable online learning and dynamic refinement of recommendations over time, ensuring adaptability to evolving fashion trends and consumer preferences.

References

1. Y. Gao, J. Li, X. Wang, and C. Zhang, "A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions," *arXiv preprint arXiv:2302.04473*, 2023.
2. X. Chen, Z. Liu, and H. Wang, "A Novel Behavior-Based Recommendation System for E-commerce," *arXiv preprint arXiv:2403.18536*, 2024.
3. A. Kwon, H. Lee, and J. Kim, "A Multimodal Recommender System Using Deep Learning Techniques Combining Review Texts and Images," in *Proc. IEEE Int. Conf. on Big Data (Big Data)*, 2021, pp. 191–198.
4. Z. Zhao, X. Liu, and L. Zhang, "A Comprehensive Survey on Multimodal Recommender Systems," *ACM Computing Surveys*, vol. 53, no. 4, pp. 190–210, 2020.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
7. S. Feng, X. Li, and L. Zhang, "Late Fusion for Multimodal Recommendation," in *Proc. 27th ACM Int. Conf. on Information and Knowledge Management (CIKM)*, 2018, pp. 1993–1996.
8. Kaggle, "Fashion Product Images (Small)," Available: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small.
9. L. Wang, H. Xu, and D. Zhang, "Multimodal Fusion for Fashion Recommendation," in *Proc. ACM Conf. on Recommender Systems*, 2018, pp. 165–173.
10. Y. Zhang, D. Xu, and Q. Liu, "Fashion Recommendation with Visual and Textual Features," in *Proc. IEEE Int. Conf. on Computer Vision*, 2018, pp. 3522–3531.
11. M. Elhoseiny, M. Saleh, and A. Elgammal, "Fashion Model: A Visual and Semantic Approach for Fashion Image and Text Matching," *IEEE Trans. on Multimedia*, vol. 19, no. 10, pp. 2323–2332, 2017.
12. X. Wei, T. Li, and Z. Zhang, "Multimodal Learning for Fashion Recommendation Systems," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2019, pp. 368–373.
13. Y. Gao, Y. Liu, and L. Wang, "FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval," *IEEE Trans. on Multimedia*, vol. 22, no. 1, pp. 180–189, 2020.
14. G. Konnurmath and S. Chickerur, "Power-Aware Characteristics of Matrix Operations on Multicores," *Journal of Intelligent and Robotic Systems*, pp. 2102–2123, Published online: Dec. 29, 2021. [Online]. Available: https://doi.org/10.1080/08839514.2021.1999013
15. G. Konnurmath and S. Chickerur, "GPU Shader Analysis and Power Optimization Model," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12925–12930, Feb. 2024. [Online]. Available: https://doi.org/10.48084/etasr.6695