Design of a Multimodal Product Recommendation System Using Image and Text Features

Ssvr Kumar Addagarla[1], Upendra Kumar[2]
[1] Lincoln University College, Malaysia ; [2] Institute of Engineering & Technology, Lucknow, India,
Adjunct research faculty, Lincoln University College, Malaysia;
Email ID : pdf.ssvrkumar@lincoln.edu.my, undera.ietlko@gmail.com

---

Abstract: The rapid growth of e-commerce platforms has increased the need for effective recommendation systems to help users discover relevant products from large online catalogs. Traditional recommender systems mainly rely on user interaction data or single-source information, which often limits their ability to capture detailed product characteristics. In the proposed framework, product images are processed using deep learning-based feature extraction models, while textual information such as product titles and descriptions is encoded using semantic embedding techniques. The extracted features are then combined to form a unified product representation, which is used to compute similarity among products and generate recommendations. The system architecture, feature extraction process, and recommendation strategy are discussed in detail. The proposed framework provides a simple foundation for developing multimodal recommendation systems in e-commerce environments. Future work will focus on implementing the system on real-world datasets and evaluating its performance using standard recommendation metrics.
Keywords: Multimodal recommendation; Product clustering; Vision Transformers; Explainable AI; E-commerce.

Introduction

The raise in expansion of e-commerce platforms has largely increased the number and various diversity of available products. The Online global marketplaces such as Amazon, Flipkart, and Myntra host millions of items across multiple categories. With such scale, machine learning based recommender systems play a vital role in helping users discover relevant products and improving customer engagement. Traditional recommendation approaches such as collaborative filtering and content-based filtering have been widely used in e-commerce platforms [1]. However, these recommender systems primarily focus on user interaction patterns or various category level purchases similarity and often fail to capture the low level fine-grained product attributes. Whereas the Modern products frequently share similar visual appearances and textual descriptions. Considering an example, two products may belong to the same category but differ in subtle attributes such as fabric type, design pattern, color tone, or usage context.

Recent studies have highlighted the importance of integrating multimodal information, such as images and text, to enhance recommendation systems [2], [3]. Multimodal machine learning enables the extraction of complementary features from different data modalities and improves the semantic understanding of products [4]. In addition to limited product understanding, most existing recommendation systems operate as black-box models. Explainable artificial intelligence (XAI) has

therefore become an important research direction, aiming to provide transparency and interpretability in recommendation systems [5].

To address these limitations, this research proposes a unified multimodal framework that integrates visual and textual product information using modern deep learning architectures. The proposed system combines fine-grained multimodal representation learning, attribute-aware clustering, compatibility-aware recommendation, and explainable AI mechanisms.

Proposed Methodology

The proposed methodology consists of four main components:

1. Multimodal Representation Learning
2. Fine-Grained Product Clustering
3. Compatibility-Aware Recommendation
4. Explainability Engine

Multimodal Representation Learning

The first stage of the system focuses on learning a unified representation of products by integrating visual and textual information. The Product images are processed using Vision Transformers (ViT), which apply the self-attention mechanisms to the image patches for capturing various relationships within the given images [6]. Further Compared with convolutional neural networks, Vision Transformers are better suited for modeling fine-grained visual patterns. Later in the process, textual product information is encoded using BERT models and a transformer-based model designed to generate semantically meaningful sentence embeddings [7]. The BERT models are effectively captures contextual relationships in product descriptions and metadata. Once the image and text features are extracted, the visual and textual embeddings are fused through a projection layer to produce a shared embedding space. Similar approaches have been successfully used in multimodal recommendation systems and vision-language models [8].

Fine-Grained Product Clustering:

After obtaining the multimodal fusion embeddings from the multimodal fusioning, clustering techniques are applied to identify latent product groupings.

The following two clustering algorithms are used:

- KMeans Clustering for partition-based clustering
- HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) for density-based clustering [9]

This Clustering mechanism allows the discovery of various latent attributes such as product style, material, or functional usage that may not be explicitly labeled. This improves the interpretability of product groupings and supports compatibility-aware recommendations.

Compatibility-Aware Recommendation:

Further in the pipeline of the process, the recommendation module retrieves candidate products and ranks them using similarity and compatibility scores.

An Efficient similarity search is performed using FAISS (Facebook AI Similarity Search) or HNSW (Hierarchical Navigable Small World graphs), which enable scalable nearest-neighbor retrieval in high-dimensional embedding spaces [10].

The final recommendation score is calculated by combining:

- Similarity Score (embedding distance)
- Complementarity Score (attribute compatibility or co-purchase patterns)

This strategy ensures that recommendations include both similar and functionally complementary products.

Explainability Engine (XAI Module):

To enhance transparency, a multi-level explainability module is incorporated. Visual Attention maps from Vision Transformers and Grad-CAM techniques are proposed to be use to highlight image regions that influence the recommendation decision [11]. Later, Token importance analysis identifies key descriptive words contributing to the recommendation. Large Language Models can be used to generate human-readable explanations based on the provided visual and textual evidence. Many Recent studies demonstrate the effectiveness of LLM-based explanation generation in recommender systems [12]. Finally Faithfulness testing is conducted using feature removal experiments to ensure that explanations accurately reflect model behavior.

Experimental Setup and Evaluation

Experiments will be conducted using publicly available multimodal datasets such as the H&M Personalized Fashion Recommendation dataset and Amazon Fashion dataset [13]. The system will be implemented using Python and PyTorch, leveraging pretrained transformer models for representation learning. The following Table 1 depicts the various evaluation metrics are to be validated during the experimentation process.

Table 1: Experimental validation metrics

Method	**Objective**	**Evaluation Metrics**
Multimodal Representation Learning	Validate effectiveness of multimodal fusion over unimodal models	• Retrieval Accuracy • Cosine Similarity Ranking • UMAP / t-SNE Visualization • Cluster Coherence Score
Fine-Grained Product Clustering	Discover latent attribute-based product groupings	• Silhouette Score • Davies–Bouldin Index • Attribute Consistency Validation
Compatibility-Aware Recommendation	Improve ranking using similarity + complementarity	• Precision@K • Recall@K

		• NDCG(Normalized Discounted Cumulative Gain)@K • MRR (Mean Reciprocal Rank)
Explainability Engine (XAI)	Ensure transparent and faithful recommendation explanations	• Explanation Faithfulness Score • Human Evaluation (Clarity, Trust) • ROUGE / BLEU

Conclusion

This research work presented a comprehensive methodology for a unified, multimodal recommendation framework. Rather than relying on a single data source, our approach merges several sophisticated layers: we use Vision Transformers to see and encode product imagery, alongside semantic text embeddings to capture the nuance of product descriptions. To ensure the results are both relevant and organized, we've integrated fine-grained clustering and a compatibility-aware ranking system that understands how products actually fit together. Beyond just making accurate suggestions, we also prioritized explainable AI (XAI) techniques. The goal here is twofold: to significantly boost the precision of e-commerce recommendations and to pull back the curtain on *why* certain products are being suggested. By doing so, we aim to create a recommendation environment that is not only more effective but also far more transparent and trustworthy for the end user.

References

1. Zhang, Y., & Chen, X., 2020. "Explainable Recommendation: A Survey and New Perspectives." *Foundations and Trends in Information Retrieval*, 14(1), 1–101.
2. Baltescu, P., Ororbia, A., Smith, S., & Pineau, J., 2017. "MRNet-Product2Vec: A Multi-task Recurrent Neural Network for Product Embeddings." *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*.
3. Zhou, K., Yang, H., Zhou, Y., & Wen, J., 2022. "CLIP4Rec: A New Vision-Language Foundation Model for Recommendation." *arXiv preprint arXiv:2209.12356*.
4. Baltrusaitis, T., Ahuja, C., & Morency, L., 2019. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.
5. Akhtar, R., Gupta, S., & Sharma, M., 2024. "Decoding the Recommender System: A Comprehensive Guide to Explainable AI in E-commerce." *Expert Systems with Applications*, 238.
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2021. "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale." *Proceedings of the International Conference on Learning Representations (ICLR)*.
7. Reimers, N., & Gurevych, I., 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3982–3992.

8. Radford, A., Kim, J., Hallacy, C., et al., 2021. “Learning Transferable Visual Models from Natural Language Supervision.” *Proceedings of the International Conference on Machine Learning (ICML)*.
9. McInnes, L., Healy, J., & Astels, S., 2017. “HDBSCAN: Hierarchical Density-Based Clustering.” *Journal of Open Source Software*, 2(11), 205.
10. Johnson, J., Douze, M., & Jégou, H., 2019. “Billion-scale Similarity Search with FAISS.” *IEEE Transactions on Big Data*, 7(3), 535–547.
11. Selvaraju, R., Cogswell, M., Das, A., et al., 2017. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.
12. Tsai, Y., Lin, H., & Chen, J., 2024. “XRec: Large Language Models for Explainable Recommendation.” *arXiv preprint arXiv:2406.02377*.
13. H&M Group, 2021. “H&M Personalized Fashion Recommendation Dataset.” *Kaggle Dataset*.