Multimodal Explainable Fashion Recommendation System

Aasmaan Gupta - IMT2021006
IIIT Bangalore
aasmaan.gupta@iiitb.ac.in

Trupal Patel - IMT2021056
IIIT Bangalore
trupalkumar.patel@iiitb.ac.in

Kevin Adesara - IMT2021070
IIIT Bangalore
kevin.adesara@iiitb.ac.in

Anant Ojha - IMT2021102
IIIT Bangalore
anant.ojha@iiitb.ac.in

I. Abstract

This report introduces a Multimodal Explainable Fashion Recommendation System that uses the Amazon Fashion Dataset, which includes product reviews and metadata. By combining textual and visual information through vector representations and features from a VGG model, the system utilizes three neural networks to handle user representation, determine image region significance, and predict ratings. Despite facing challenges with computational power and dataset size, our approach tries to captures user preferences and provides personalized recommendations. We also examine the system's performance, accuracy of predictions, and qualitative aspects, showcasing the strengths and challenges of using multimodal recommendation systems in the online fashion retail sector.

II. Introduction

E-commerce has revolutionized the fashion retail industry, driving the need for sophisticated recommendation systems that offer personalized shopping experiences. Traditional systems often depend solely on textual or visual data, which can restrict their effectiveness. Customers today expect more tailored and intuitive recommendations that reflect their individual tastes and preferences. This project aims to create a Multimodal Explainable Fashion Recommendation System that merges textual reviews and visual features to provide more accurate and personalized fashion suggestions. By leveraging the Amazon Fashion Dataset[1], our system integrates and analyzes these multimodal inputs through advanced neural networks[2], capturing a richer understanding of user preferences and item characteristics.

In developing this system, we employ three neural networks: one for user representation, another for determining the significance of different image regions, and a third for predicting ratings. These networks work together to generate recommendations that are not only accurate but also explainable, helping users understand why certain products are suggested to them. This report covers the preprocessing steps required to prepare the dataset, the detailed architecture of the recommendation algorithm, and a thorough evaluation of its performance. We highlight the system's strengths, such as its ability to handle complex user preferences, as well as its limitations, including computational constraints and the need for larger datasets. Our findings offer valuable insights into the potential and challenges of implementing multimodal recommendation systems in the fashion e-commerce landscape.

III. Dataset

A. *Dataset Description*

We utilized the Amazon Fashion Dataset to develop our Fashion Recommendation System. This dataset comprises product reviews and metadata from Amazon's Fashion category, offering valuable insights into products, reviewers, ratings, and review text.

Review Data:

- The review data encompasses 883,636 reviews contributed by various reviewers.
- It includes crucial information such as reviewer ID, product ID, review text, rating, summary, and review time.

Metadata:

- The metadata provides supplementary details about the products.
- It encompasses product ID, title, price, brand, and categories.

B. *Dataset Analysis and Preprocessing*

We begin by analyzing the dataset to understand its characteristics. The Amazon Fashion Dataset serves as the foundation for our Fashion Recommendation System. This dataset comprises reviews from various reviewers, encompassing a wealth of information about products, reviewers, ratings, and review text.

*1) **First Steps**:* To gain insights into the dataset, we conducted an initial analysis. Plot 1 illustrates the distribution of the number of users who voted for less than x number of items. From the plot, it is evident that the number of users who voted less than x items reaches its peak for x around 35.

Given the size of the dataset, we opt to use a subset of the data for model training.

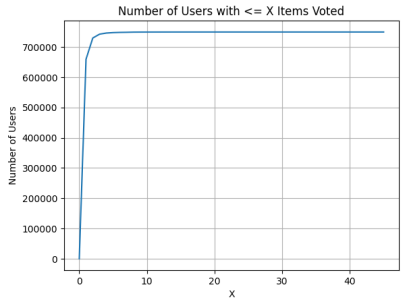

Fig. 1. Number of users who voted less than x items

*2) **Extracting Subset**:* We proceed by extracting a subset of the dataset to train our model. We create a new table, as shown in Figure 2, consisting of relevant columns such as User ID, Item (product ID), Review, overall rating, and other metadata. Notably, we sort the entries in decreasing order of the number of items voted for each review item. Subsequently, our model will be trained on the top 10,000 entries from this sorted list.

User ID	Item ID	ImageURL	Detail	Rating	Review

Fig. 2. Columns of the dataset

These initial steps in dataset analysis and preprocessing lay the groundwork for our Fashion Recommendation System. Further preprocessing steps will be discussed in subsequent sections.

*3) **Extracting Image Representations**:* To incorporate visual information into our Fashion Recommendation System, we extract image representations from the product images. We employ a VGG [3] model to process the images and extract features. Considering that users may have diverse areas of interest within the images, we divide each image into four windows and extract representations for each window.

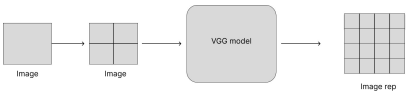

Fig. 3. Image Representations

Figure 3 showcases the image representations obtained from the VGG model. By leveraging the VGG model, we extract rich image features that capture essential visual characteristics, such as colors, textures, and patterns, enabling our recommendation system to incorporate visual information into the recommendation process effectively.

4) Extracting Vector Representations for Texts: In our Fashion Recommendation System, textual information plays a crucial role in understanding user preferences and product characteristics. To represent textual data effectively, we utilized encoders[4] to obtain vector representations for review text, item descriptions, and summaries.

We made a deliberate choice not to use pretrained encoders to ensure that the vector representations are specifically tailored to fashion recommendation. By training the encoders on our dataset, we aimed to capture domain-specific features and nuances that are relevant to fashion products. Please note that the Item Detail attribute is created by concatenating various relevant attributes given in the meta data of the item.

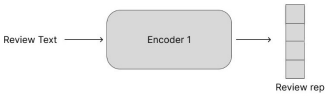

Fig. 4. Review Text Representations

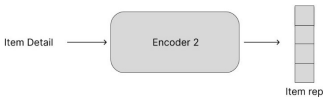

Fig. 5. Item Description Representations

Figures 4 and 5 depict the vector representations obtained for review text and item descriptions, respectively. These representations encode semantic information and enable effective modeling of user-item interactions in our recommendation system.

IV. Recommendation Algorithm

1. User Representation (First Neural Network)

The first neural network serves the crucial task of obtaining the final user representation by integrating information from both textual and visual modalities. It takes as inputs the vector representations obtained from the encoders for the user's review text and item description, as well as the feature representations of subimages generated by the VGG model. These inputs are then processed to derive a comprehensive user representation that encapsulates both textual preferences expressed in reviews and visual preferences inferred from the images. By synthesizing information from multiple modalities, the first neural network captures a holistic view of the user's preferences, laying the foundation for personalized recommendation generation.

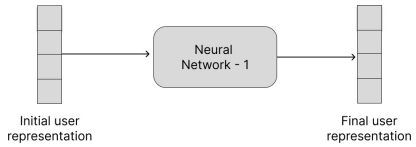

Fig. 6. User Representation

2. Alpha Calculation (Second Neural Network)

The second neural network plays a pivotal role in determining the significance of different regions within the images for recommendation purposes. It takes as inputs the feature representations of subimages obtained from the VGG model, the final user representation derived from the first neural network, and the item representation. Utilizing this information, the second neural network computes weights, referred to as alphas, which indicate the relative importance of each subimage in influencing the user's preferences. These alphas serve as attention mechanisms, dynamically adjusting the contribution of each subimage based on its relevance to

the user's preferences, thereby enhancing the discriminative power of our recommendation system.

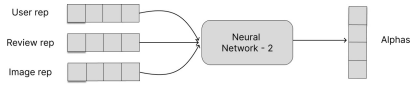

Fig. 7. Alphas

3. Rating Prediction (Third Neural Network)

The third neural network is responsible for predicting ratings for the items based on the integrated information from both textual and visual modalities. It takes as inputs the weighted image representation, derived by combining the subimage representations using the alphas, along with the final user representation and the item representation. By synthesizing these representations, the third neural network generates predictions for the ratings of the items, thus providing personalized recommendations tailored to the user's preferences. Through the collaborative efforts of the three neural networks and the supporting encoders and VGG model, our recommendation algorithm achieves enhanced performance and user satisfaction, effectively capturing the multidimensional nature of user preferences in the fashion domain.

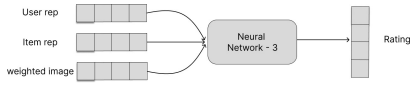

Fig. 8. Ratings

The Complete Model Diagram)

V. Inference

The final recommendation is done as follows:

1) Input text from the user specifying their preferences is received.
2) The input text is fed into our neural network to obtain ratings for all the products in the dataset.
3) From the top-rated products, a subset is randomly selected, and their images are displayed to the user.

This process ensures that users receive personalized recommendations based on their preferences, as inferred from the input text provided. It enhances user engagement and satisfaction with our recommendation system.

For completely new user we select random items to show the users until atleast one purchase is done.

VI. Analysis

In this section, we analyze various aspects of our Fashion Recommendation System, including computational performance, prediction accuracy, and qualitative evaluation.

A. Computational Performance

The computational performance of our system is an important consideration, especially given the size of the dataset and the complexity of the models involved. We found that the total running time for our code, using only 10,000 reviews and vector representations of dimension 16x1, took approximately 7 to 8 hours to complete. This highlights the computational challenges associated with training and testing large-scale recommendation systems.

B. Prediction Accuracy

One key aspect of our analysis involves examining the predicted ratings compared to the actual ratings for user-item pairs. Upon analyzing the results, we observed significant variations between predicted and actual ratings. This discrepancy could be attributed to several factors:

- **Data Subset**: We trained our model on a much smaller subset of the dataset (10,000 reviews), which may not capture the full diversity and complexity of user-item interactions present in the entire dataset.
- **Model Complexity**: The simplicity of our models, such as the use of shallow neural networks and limited training data, may not sufficiently capture the underlying patterns in the data.
- **Feature Representation**: The vector representations obtained for textual data and images may not fully capture the semantic and visual characteristics relevant to fashion recommendation.

To visualize the performance of our recommendation system, we generated a plot comparing predicted ratings to actual ratings for a subset of user-item pairs. Figure 10 illustrates this comparison.

C. Qualitative Evaluation

In addition to quantitative analysis, we conducted qualitative evaluation to assess the effectiveness of our recommendation system. One approach involved highlighting portions of images based on the review text using alpha values obtained from the trained neural network. As an example, we considered an image of heels and highlighted the regions with higher alpha values to emphasize relevant features.

Figures 11 and 12 depict the original image of heels and the corresponding highlighted image, respectively. This qualitative evaluation provides insights into how our recommendation system leverages both textual and visual information to make personalized recommendations.

VII. Limitations

While our recommendation system demonstrates promising performance, it is important to acknowledge several limitations that may impact its effectiveness and applicability:

- Computational Constraints: Due to computational limitations, our encoder layers and representation sizes are kept minimal. This may restrict the system's ability to capture intricate user preferences and item features comprehensively.

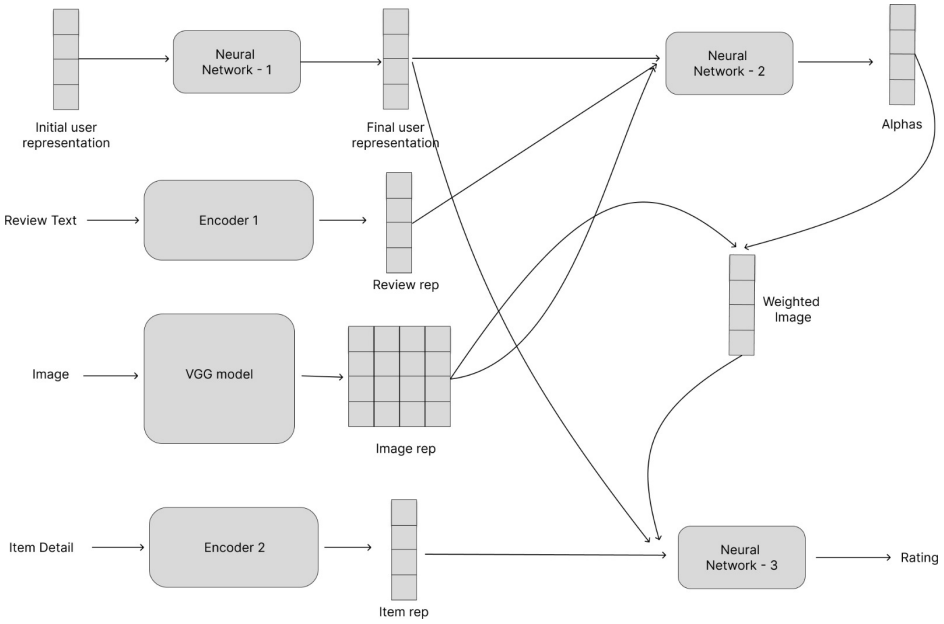

Fig. 9. The complete working of full recommendation process

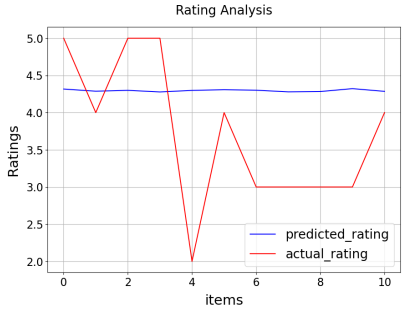

Fig. 10. Comparison of Predicted and Actual Ratings

Fig. 11. Original Image: Heels

Fig. 12. Highlighted Image: Heels

- Dataset Size: Our models are trained on a relatively small dataset comprising approximately ten thousand data entries due to computational complexity. Limited training data may lead to suboptimal model generalization and may not fully capture the diversity of user preferences and item characteristics present in larger datasets.
- Simplified Image Regions: To mitigate computational complexity, image regions are divided into only four segments instead of more granular divisions. This simplification may limit the system's ability to capture fine-grained visual features and nuances, potentially affecting the accuracy of image-based recommendations.

Acknowledging these limitations provides valuable insights for future research and development efforts. Addressing these constraints could involve exploring more efficient computational strategies, acquiring larger and more diverse datasets for training, and refining the granularity of image regions to enhance recommendation accuracy and robustness.

VIII. Conclusion

In conclusion, our Multimodal Explainable Fashion Recommendation System represents a significant step towards enhancing personalized shopping experiences in the fashion retail industry. By integrating textual reviews and visual product features from the Amazon Fashion Dataset and leveraging advanced neural networks for user representation, image region significance, and rating prediction, our system aims to provide more accurate and tailored recommendations. Despite facing challenges such as computational constraints, a limited dataset size, and suboptimal results, this project highlights the potential of multimodal systems in e-commerce. Our findings provide valuable insights and a solid foundation for future improvements in computational efficiency, dataset expansion, and model refinement to better meet the evolving demands of the fashion retail market.

References

[1] Julian McAuley. *Amazon Product Data (2018)*. https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/. Accessed on 2024-05-15.

[2] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural networks* 61 (2015), pp. 85–117.

[3] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[4] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).