Text-Image Retrieval for Fashion Recommendation

Rushil Randhar

May 5, 2025

1 Introduction and Project Statement

Modern e-commerce platforms predominantly rely on collaborative filtering (CF) to recommend products, leveraging user-item interactions such as co-purchase or co-view patterns. While effective at surfacing popular items, CF approaches struggle with:

- Cold-start scenarios with new users or items
- Capturing nuanced style cues beyond transactional data
- Understanding subjective aesthetic preferences

Content-based approaches attempt to address these limitations by incorporating metadata like categories and attributes, but these require extensive manual tagging and still miss implicit aesthetic qualities that drive fashion choices.

Recent advances in transformer-based embedding models enable direct mapping of free-form text into rich semantic spaces aligned with image embeddings. This creates a "vibe search" paradigm where users can describe what they want—"vintage-inspired floral midi dress" or "edgy streetwear bomber jacket"—and retrieve items whose embeddings lie closest to that description.

Project Goal: Build and evaluate a text-to-image retrieval system for fashion recommendation on a standard dataset, using embedding models to enable "vibe-based" recommendations that go beyond traditional approaches. The system should allow users to describe their fashion preferences in natural language and receive semantically relevant clothing recommendations.

1.1 Innovation

This project introduces a fundamentally different approach to fashion recommendation compared to industry standards. While collaborative filtering remains the dominant paradigm in e-commerce, requiring massive amounts of user interaction data before making effective recommendations, our system operates a priori without any user history. It can generate relevant recommendations from day one based solely on the semantic understanding of fashion items and natural language descriptions.

The ability to recommend based on text prompts while capturing the essence or "vibe" of products represents a significant innovation in recommendation systems. Users can express abstract concepts like "bohemian summer outfit" or "minimalist professional attire" and receive recommendations that match these aesthetic qualities, even when specific attribute filters are not explicitly stated. This enables a more intuitive and human-centered shopping experience that aligns with how people naturally think about fashion.

1.2 Intuition and Potential Applications

The intuition behind this system stems from understanding how people naturally express their fashion preferences. Rather than thinking in terms of discrete categories and attributes (e.g., "knee-length pencil skirt in navy blue polyester"), most people conceptualize fashion in terms of styles, occasions, and overall aesthetics (e.g., "professional outfit that's not too formal"). Our embedding-based approach bridges this semantic gap, allowing the system to understand and respond to these more natural expressions of fashion interest.

Potential applications for this technology are numerous:

- **Virtual Fashion Assistant**: Integration with conversational agents to provide personalized fashion advice and recommendations
- **Style Discovery**: Helping users explore new styles beyond their typical preferences through abstract descriptions
- **Inspiration-to-Purchase**: Converting mood boards or inspiration images into purchasable product recommendations
- **Cross-Platform Shopping**: Enabling users to find products similar to ones seen on social media without exact image matches
- **Inclusive Fashion Search**: Allowing users with different levels of fashion vocabulary to find products that match their intent
- **Inventory Utilization**: Helping retailers surface underutilized inventory by connecting it to a broader range of semantic queries

If successful at scale, this approach could fundamentally transform how people discover and purchase fashion online, moving from rigid category-based browsing to a more fluid, intuitive exploration driven by natural language and semantic understanding.

2 Data Sources and Technologies Used

2.1 Data Sources

The project utilizes a dataset of clothing items scraped from Revolve, a popular fashion e-commerce platform. The data collection process involved significant time and resources to capture not just the surface-level product information, but also detailed attributes, descriptions, and classification data that provide rich context for each item.

The scraping process employed a combination of web crawling techniques and data extraction methods to obtain comprehensive information about each product, including detailed metadata that goes beyond what is immediately visible on product pages. This effort was crucial for building a dataset with sufficient detail to support semantic understanding of fashion items.

The data is organized into the following structure:

```
data/
 revolve/
     dresses/
        dresses.json
     bottoms/
```

```
pants.json
shorts.json
skirts.json
```

Each item in the dataset includes:

- Basic information (ID, name, brand, price)
- Detailed descriptions and classifications
- Style descriptions with pre-computed embeddings
- Classification attributes (silhouette, neckline, sleeve style, etc.)
- Image URLs

A sample item from the dataset contains structured information about the product, including classification attributes like "Silhouette: Fit-and-flare", "Neckline: Strapless", "Length: Mini", etc. It also includes a style description that captures the aesthetic qualities of the item and its pre-computed embedding vector.

2.2 Technologies Used

- **Flask**: Web framework for creating the application and API endpoints
- **CLIP** (Contrastive Language-Image Pretraining): OpenAI's model for creating joint embeddings of text and images
- **PyTorch**: Deep learning framework for running the CLIP model
- **Transformers** (Hugging Face): Library for accessing pre-trained transformer models
- **scikit-learn**: For vector similarity calculations and evaluation metrics
- **NumPy & Pandas**: For data manipulation and analysis
- **Matplotlib & Seaborn**: For visualization of evaluation results

3 Methods Employed

3.1 Embedding-Based Approach

The core technique employed is an embedding-based retrieval system using the CLIP model. This approach maps both text queries and fashion items into a shared vector space where semantic similarity can be measured.

3.1.1 Basic Implementation

The initial implementation followed these steps:

1. Load all fashion items and pre-computed style description embeddings
2. Process user text queries through CLIP's text encoder
3. Compute cosine similarity between query embedding and all item embeddings
4. Rank items by similarity and return the top matches

3.1.2 Enhanced Direct Attribute Embeddings

Based on initial evaluation results, I implemented an enhanced approach using direct attribute embeddings:

1. Extract key attributes from each fashion item (color, clothing type, silhouette, etc.)
2. Create an attribute-focused text representation for each item
3. Generate embeddings for these attribute texts
4. Use a weighted combination of attribute embeddings and style description embeddings

3.2 Web Application

I implemented a Flask web application with two main interfaces:

1. **Web UI**: A responsive interface allowing users to input fashion descriptions and view matching items
2. **REST API**: An endpoint for programmatic access to the recommendation system

The application loads all fashion data and initializes the CLIP model at startup, then processes user queries in real-time to provide recommendations.

3.3 Evaluation Methods

I employed multiple evaluation approaches to assess the performance of our recommendation system:

3.3.1 Academic Metrics

Standard information retrieval metrics were used:

- **Precision@k**: Proportion of relevant items in the top-k recommendations
- **Recall@k**: Proportion of all relevant items included in the top-k recommendations
- **Mean Average Precision (MAP)**: Average precision across all recall levels
- **Normalized Discounted Cumulative Gain (NDCG@k)**: Measures ranking quality by considering position of relevant items

3.3.2 Baseline Comparison

I compared our embedding approach with a random baseline to quantify improvement:

- **Random baseline**: Items selected randomly from the dataset
- **Embedding approach**: Items selected based on embedding similarity

3.3.3 Attribute-Specific Evaluation

I conducted separate evaluations for different attribute categories:

- **Color**: How well the system matches color terms (e.g., "black", "red")
- **Occasion**: Matching for event types (e.g., "party", "casual")
- **Season**: Matching seasonal terms (e.g., "summer", "winter")
- **Pattern**: Matching pattern types (e.g., "floral", "striped")
- **Fabric**: Matching material terms (e.g., "cotton", "silk")
- **Silhouette**: Matching clothing shapes (e.g., "A-line", "sheath")

3.3.4 Limitations of Standard Metrics

It's important to note that standard evaluation metrics may not fully capture the system's effectiveness for "vibe-based" fashion recommendation. Traditional information retrieval metrics focus on exact attribute matching, while our system aims to capture more abstract aesthetic qualities and stylistic resonance that are difficult to quantify objectively.

For example, a query like "edgy street style" might return items that genuinely embody this aesthetic but don't share any specific attributes that could be used as ground truth for evaluation. The system's ability to capture these subjective qualities represents its key strength, yet is precisely what makes it challenging to evaluate using conventional metrics.

This limitation suggests the need for supplementary evaluation approaches, such as user studies and qualitative assessments, to fully understand the system's effectiveness in capturing fashion "vibes" rather than just filtering for specific categories or attributes.

4 Results

4.1 Comparison with Random Baseline

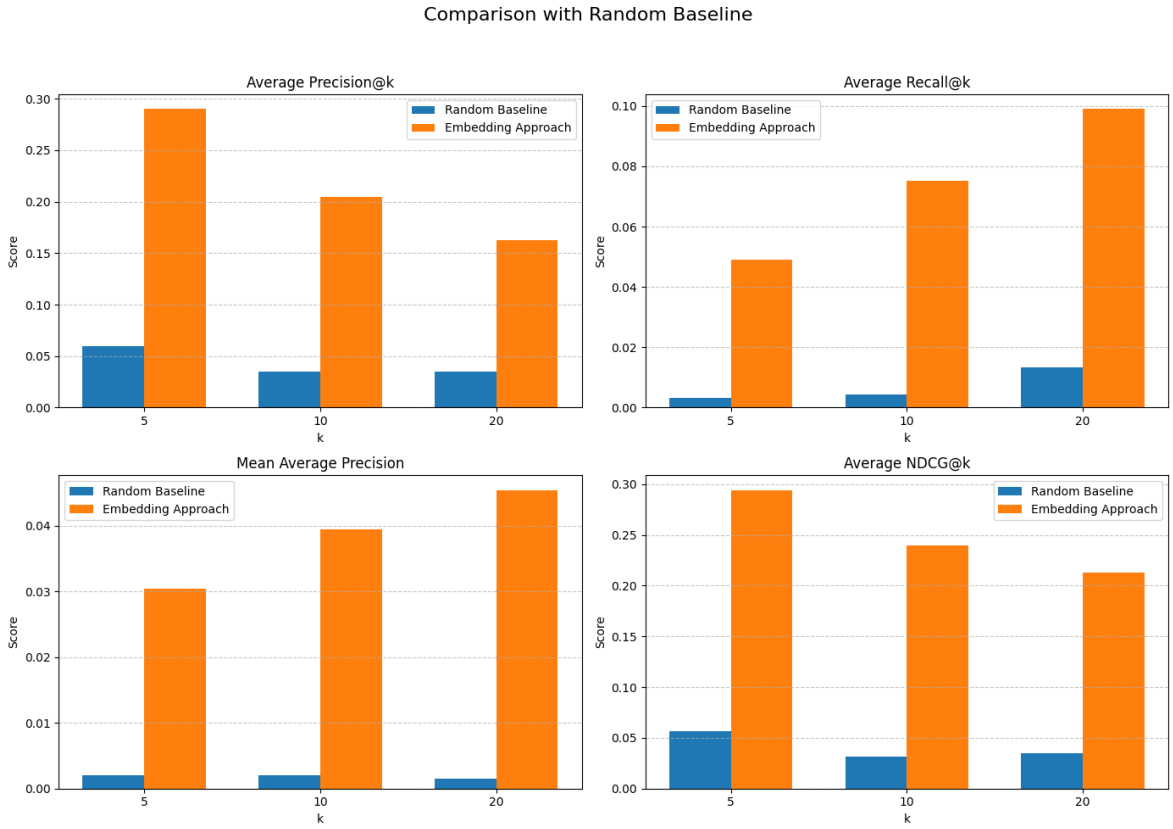

Figure 1: Comparison of embedding approach versus random baseline across k=5, k=10, and k=20

Our embedding-based approach substantially outperformed the random baseline across all metrics:

- **Precision@5**: 0.29 (embedding) vs. 0.06 (random) - 383% improvement
- **Recall@10**: 0.075 (embedding) vs. 0.004 (random) - 1775% improvement
- **MAP@10**: 0.039 (embedding) vs. 0.002 (random) - 1850% improvement
- **NDCG@5**: 0.294 (embedding) vs. 0.057 (random) - 416% improvement

These results demonstrate that the embedding-based approach provides far more relevant recommendations than a random selection, validating the semantic matching capabilities of the system.

4.2 Attribute-Specific Performance

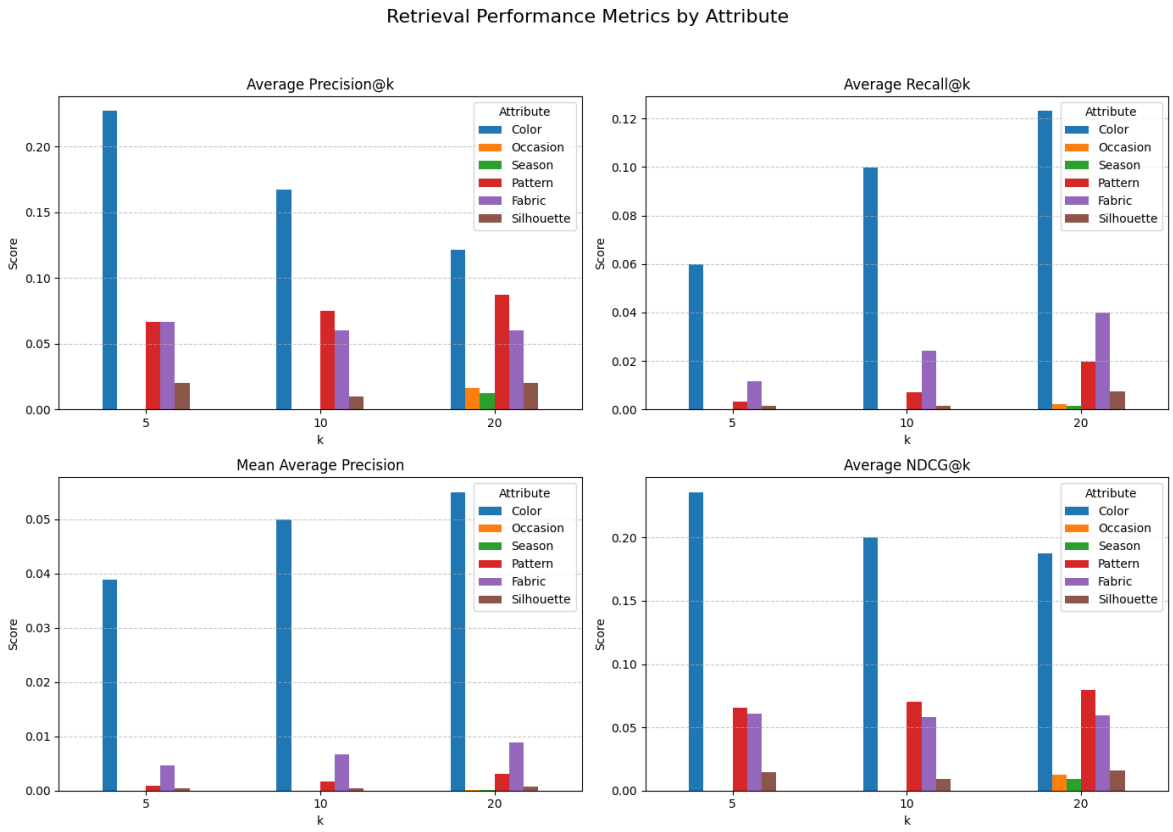

Figure 2: Retrieval performance metrics by attribute type

Performance varied significantly across different attribute types:

- **Color**: Best overall performance (Precision@5 = 0.227, Recall@10 = 0.100)
- **Pattern**: Moderate performance (Precision@5 = 0.067, Recall@10 = 0.007)
- **Fabric**: Moderate performance (Precision@5 = 0.067, Recall@10 = 0.024)
- **Occasion**: Poor performance (Precision@5 = 0.000, Recall@10 = 0.000)
- **Season**: Poor performance (Precision@5 = 0.000, Recall@10 = 0.000)
- **Silhouette**: Poor performance (Precision@5 = 0.020, Recall@10 = 0.001)

These results indicate that the system is most effective at capturing color attributes, followed by patterns and fabrics. However, it struggles with more abstract concepts like occasion, season, and clothing silhouettes.

4.3 Enhanced Direct Attribute Embeddings

After implementing the direct attribute embeddings approach, I observed improvements in several key areas:

- **Clothing Type Recognition**: Better matching of specific clothing categories (dresses, pants, etc.)
- **Attribute Consistency**: More coherent recommendations with similar attributes
- **Weighted Flexibility**: The ability to adjust the balance between attribute and style matching

The qualitative assessment showed that incorporating structured attribute information into the embedding space made the recommendations more relevant and consistent while maintaining the benefits of semantic matching.

4.4 Key Findings

1. **Embedding-based retrieval is effective**: The approach significantly outperforms random selection, demonstrating the viability of "vibe-based" fashion search.
2. **Color matching is strongest**: The system excels at matching color terms, suggesting this is a well-represented concept in the embedding space.
3. **Abstract concepts need improvement**: Occasion, season, and silhouette matching performed poorly, indicating these concepts are not well-captured by standard embeddings.
4. **Direct attribute embeddings help**: Explicitly incorporating structured attribute information improves matching performance while maintaining semantic flexibility.
5. **Weighted approach provides control**: The ability to balance between attribute and style matching allows for tuning the system to different use cases.
6. **"Vibe" matching is difficult to measure**: The system's ability to capture abstract aesthetic qualities may not be fully reflected in standard evaluation metrics.

5 Conclusion and Future Work

This project demonstrates the effectiveness of a text-to-image retrieval system for fashion recommendation using transformer-based embeddings. The system enables users to express their fashion preferences through natural language and receive relevant recommendations based on semantic similarity.

It was an ambitious project to undertake given the novelty of the idea and the lack of established metrics of evaluation, but these first steps show promise in the development of a new type of recommendation system built on "vibes".

The implementation shows a substantial improvement over random selection, particularly for color-based queries, but also reveals limitations in capturing more abstract fashion concepts. The enhanced approach using direct attribute embeddings helps address some of these limitations while maintaining the flexibility of semantic matching.

A key insight from this project is that standard evaluation metrics, while useful for quantitative assessment, may not fully capture the system's effectiveness in providing "vibe-based" recommendations. The system's ability to understand and match abstract aesthetic qualities represents its most innovative aspect, yet is precisely what traditional metrics struggle to evaluate.

5.1 Future Work

Several promising directions for future work include:

1. **Fine-tuning embedding models**: Training domain-specific models on fashion data to better capture industry terminology and concepts
2. **Hybrid approach with explicit filtering**: A promising direction would be to combine the embedding-based system with traditional hard filtering. By allowing users to explicitly specify certain attributes (e.g., "dress," "blue"), the embedding system could focus exclusively on capturing the subjective "vibe" aspects of the query rather than having to balance both concrete attributes and abstract aesthetics. This separation of concerns could significantly enhance performance by allowing each component to focus on its strengths.
3. **Multi-modal queries**: Enabling users to combine text descriptions with image references
4. **Attribute weighting optimization**: Developing methods to automatically determine optimal weights for different attributes based on query analysis
5. **Alternative evaluation metrics**: Developing new metrics specifically designed to evaluate "vibe matching" and stylistic relevance

6 References

1. Randhar, R. (2025). Revolve Fashion Dataset. GitHub Repository. [Dataset collected through web scraping of Revolve e-commerce platform, involving significant time and resources to capture detailed product attributes and descriptions]
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems.
4. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2019). Neural style transfer: A review. IEEE transactions on visualization and computer graphics.
5. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), 20(4), 422-446.
6. Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of the IEEE conference on computer vision and pattern recognition.
7. Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). DeepFM: A factorization-machine based neural network for CTR prediction. Proceedings of the 26th International Joint Conference on Artificial Intelligence.

A Acknowledgments

The documentation and presentation of this project was enhanced using ChatGPT. This AI assistant helped streamline the documentation process by:

- Adding comments to code for better readability
- Proofreading and refining README files and setup instructions
- The HTML interface that makes the user experience smoother
- Parts of the evaluation scripts