One size does not fit all: Improving Fashion Recommendation relevance via fit-aware Neural Re-ranking

Vishal G* Sai Sreenivas Chintha* Harshitha Kolukuluru*

UMass Amherst UMass Amherst UMass Amherst

vishalg@umass.edu saisreenivas@umass.edu hkolukuluru@umass.edu

Abstract

Fashion recommendation systems play an important role in improving user experience on e-commerce platforms by providing tailored product suggestions. However, traditional recommendation models often miss the details of personal fit and style, lowering their accuracy. In this project, we aim to introduce a fit-aware neural re-ranking approach to improve fashion recommendations by incorporating custom-extracted features such as bounding box dimensions, sleeve lengths, and color attributes. Using a combination of CLIP fine-tuning and a re-ranking neural network trained on custom feature embeddings, we aim to maximize the precision and recall of the retrieved items.

1. Introduction

Image retrieval is a fundamental task in the domain of information retrieval, where given a textual description of an image, we try to retrieve the top-k images that best match this description. As e-commerce, particularly in fast fashion, has grown significantly, image retrieval systems have become integral to enhancing user experiences by delivering personalized and relevant fashion recommendations. By tailoring product suggestions to individual preferences, these systems not only improve user satisfaction but also boost sales through increased engagement by making product discovery both relevant and efficient.

Conventional approaches primarily rely on visual features and text embeddings without sufficient attention to individual fit and style preferences, which limits the precision of recommendations. This gap in precision becomes particularly relevant in the top-k recommendations, where users expect highly relevant options. In addition to this, traditional image retrieval models in e-commerce which might perform well in retrieving relevant images with high recall, often face limitations in precision (precision-recall tradeoff).

In this project, we aim to explore methods to improve the relevance of image retrieval for fashion recommendations through a fit-aware neural re-ranking approach. Specifically, we propose a two-step neural re-ranking framework that combines an initial retrieval phase using CLIP embeddings and a re-ranking phase that leverages custom features extracted from images such as bounding box dimensions, sleeve lengths, color histograms, and other style elements. Our approach aims to improve the precision by focusing on aligning items with the detailed requirements of the query. We will use datasets like Fashion30K[1] and DeepFashion[6], to retrieve a broad set of relevant items with high recall, then re-rank the top-k results by incorporating features such as bounding box dimensions, color, and sleeve length, aiming to enhance relevance (figure 1).

Figure 1. Expected improvement: In the current model (top), the ground truth image (green) is retrieved at a lower rank, whereas our proposed model (bottom) aims to increase precision by ranking the ground truth image as the top result.

2. Related Work

Cross Domain Retrieval Methods: ViLBERT[7] combines image and text embeddings by concatenating them for input into a single BERT-style encoder by jointly processing image and text. These models are pre-trained on tasks with Masked Language Modeling (MLM), allowing the model

[1] all authors contributed equally

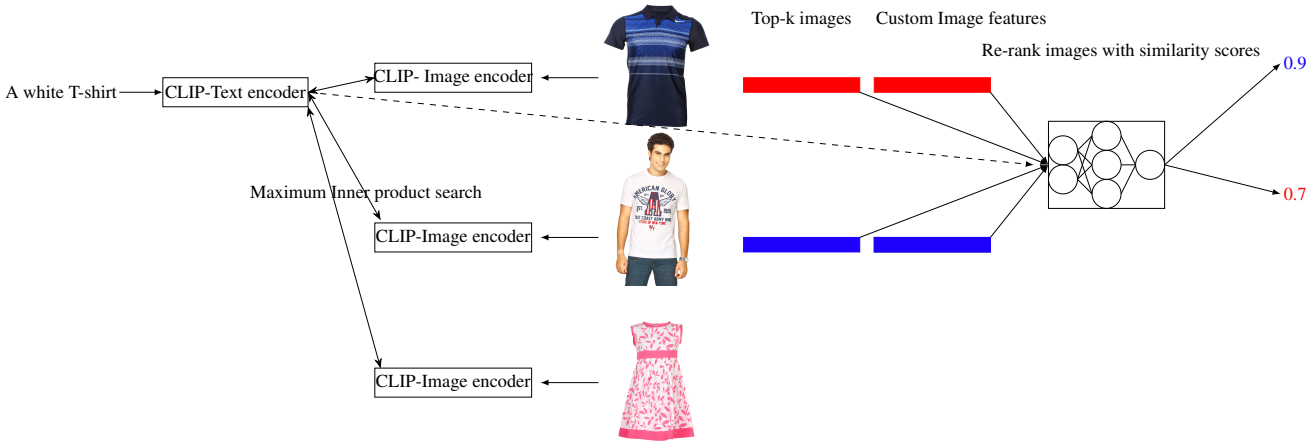

Figure 2. Diagram of our pipeline to first retrieve images with high recall and use re-ranking with custom image features such as bounding box, sleeve length etc. along with the embeddings of original image and query to achieve better precision

to build a combined understanding of visual and textual information. Some models, like UNITER[2] and OSCAR[5], include additional pre-training steps to improve alignment between image regions and text by using object categories or leveraging image depth features rather than predefined regions of interest. While effective at building a comprehensive understanding, these methods are often computationally intense and less efficient.

CLIP[8] (Contrastive Language-Image Pre-Training) aligns image and text embeddings, trained on 400 million image-text pairs. CLIP uses contrastive learning to learn joint representations of images and text. Positive pairs are taken from aligned text-image pairs while negative pairs are sampled randomly, with the InfoNCE loss[11] being optimized to align these representations via contrastive learning. A top-k item list is retrieved to maximize recall followed by a re-ranking step to optimize precision[10], where a neural network is trained to predict a relevance score $score_\theta(q, d)$ given a query q and document d, where parameters θ are learnable. This contrastive learning enables CLIP to perform zero-shot classification and retrieval tasks with competitive performance across domains, making it highly adaptable for specific applications, including fashion.

3. Technical Approach

Previous works use CLIP embeddings of texts and images to compute maximum inner product search for retrieving images given text and vice-versa. We aim to extend this pipeline by re-ranking the retrieved images using the following methods:

1. **CLIP (Baseline):** We utilized the CLIP model to generate embeddings for both images and textual queries. Through a thorough analysis of the Fashion30K[1] dataset, we identified relevant textual attributes that facilitated meaningful embedding generation. For this phase of the project, we selected a sample of 5,000 images from the dataset and computed cosine similarity between these embeddings to retrieve images for each query. We evaluated the model's performance, with results presented in the sections below, providing insights into its effectiveness and highlighting areas for improvement. The results from these evaluations will serve as a foundation for the upcoming enhancements, which aim to further refine the retrieval process.
2. **Fine-tune CLIP with custom extracted features:** We will extract custom features from images, such as bounding box dimensions using image segmentation algorithms, dominant colors using color histograms, texture features and item attributes like sleeve length, neckline type etc.
   - **Bounding Box Dimensions:** Using image segmentation algorithms like YOLO[9], Mask R-CNN[3], we will obtain the dimensions of fashion items. This helps understand their size and scale, ensuring that recommendations align with user preferences regarding fit and proportion.
   - **Dominant Colors:** We will analyze color histograms to identify the most prevalent colors in the fashion items. Understanding dominant colors is crucial for visual similarity assessments, as users often prefer items that match or complement their existing wardrobe.
   - **Texture Features:** Detailed texture characteristics can be extracted using techniques such as Local Binary Patterns (LBP) and Gabor filters. These methods will provide insights into fabric properties, which are important for consumers who have specific material pref-

erences.

- **Fashion Attributes:** We will identify stylistic elements such as sleeve length, neckline type, and pattern type through object detection and attribute classification models. By integrating these attributes, our system can offer more personalized suggestions aligned with users' individual styles, improving overall satisfaction.We will conduct ablation studies to assess the contribution of these features to re-ranking. We will fine-tune CLIP with these additional features using contrastive learning with **Low-Rank Adaptation (LoRA)** [4].

3. **Build a neural network for re-ranking:** Instead of fine-tuning CLIP with custom features, we plan to train a neural network to predict relevance scores given image embedding, query embedding and custom image features extracted in the previous step (figure 2).

Figure 3. Top 10 retrieved images for the query (red boxes indicate incorrect samples). The ground truth image (green) appears within the top 10 results.

Figure 4. Top 10 retrieved images for the query (red boxes indicate incorrect samples). The ground truth image (green) appears at rank 1.

Figure 5. Top 10 retrieved images for the query (red boxes indicate incorrect samples). The ground truth image (green) does not appear within the top 10 results.

4. Intermediate/Preliminary Results

As mentioned, we conducted experimentation with various data attributes from the Fashion30K dataset to identify effective queries for the CLIP model, ultimately generating embeddings for both text and images. With these embeddings we performed inference and computed the evaluation metrics:

1. **Quantitative Evaluation:** We computed the cosine similarity between the generated embeddings and evaluated the model's performance by measuring Recall at various values of k (R@k) 10, 30, 50. After selecting a refined subset of attributes from the text dataset, our Recall@10 improved from 32% to approximately 41%. Recall metrics were calculated at different k values, as shown in (figure 6).
2. **Qualitative Evaluation:** We observed that our baseline model was able to achieve higher recall for queries with broad attributes, often retrieving the ground truth image as the top recommendation (figure 4). However, for more nuanced queries, such as the request for a V-neck red innerwear T-shirt illustrated in (figure 3), the ground truth image was ranked 5th, while all the other images displayed different neck styles (i.e., not V-neck). Additionally, as shown in (figure 5), there were cases where the model failed to retrieve the ground truth image within the top-k recommendations, highlighting an area for improvement. This suggests that while the model excels in recognizing clear attributes, it struggles with intricate features resulting in low recall values for some examples. Next we will address these scenarios by fine-tuning CLIP on fashion specific datasets and training custom neural-network for re-ranking.

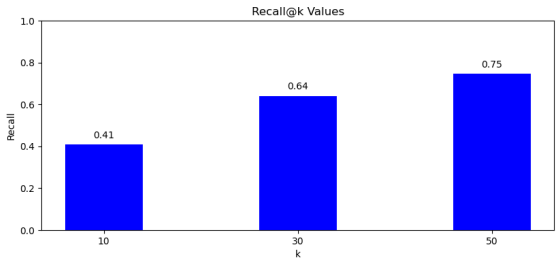

Figure 6. **Baseline:** Recall@k values for different top-k thresholds (k = 10, 30, 50) showing retrieval performance for CLIP (baseline) model.

References

[1] Param Aggarwal. Fashion product images dataset, 2019.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[5] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020.

[6] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[10] Xi SHEN, Yang Xiao, Shell Xu Hu, Othman Sbai, and Mathieu Aubry. Re-ranking for image retrieval and transductive few-shot classification. In *Advances in Neural Information Processing Systems*, pages 25932–25943. Curran Associates, Inc., 2021.

[11] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.