Prompt-Based Fashion Outfits Retrieval and Recommender System Using Binary Hashing*

Quocdung Nguyen, Hoangnam Pham, Duyhung Dao, Quangmanh Do and Vanha Tran*

FPT University, Hanoi 155514, Vietnam

Abstract

The exponential growth of e-commerce in recent years has transformed the fashion industry, propelling it into a new era of digital retail. With the convenience of online shopping, consumers now have access to an extensive array of fashion products from the comfort of their homes and as a result in need of more efficient and personalized shopping experiences. This demand paved the way for the advancement of recommendation and retrieval systems in fashion e-commerce. In this paper, we build a system plan to streamline and enhance the retrieval of fashion outfits from vast and diverse collections. Our system consists of two components, a CLIP-like model to retrieve image items matching a textual description, and a network utilizing hashing modules for efficient personalized fashion outfit recommendations. Through extensive experimentation and evaluation, we demonstrate the effectiveness of our system in providing accurate and personalized fashion outfit recommendations with desired descriptions by the consumers, like a particular color, style, occasion, season, and many more.

Keywords

Fashion Retrieval, Outfit Recommendation, Representation Learning, Hashing

1. Introduction

The fashion industry, with its ever-evolving trends and creative expressions, is a dynamic landscape characterized by ever-changing trends, styles, and personal preferences. The field has traditionally been driven by the instincts and intuitions of designers, fashion houses, and trendsetters. However, the advent of machine learning has introduced a new dimension, one where data-driven insights and algorithms wield significant influence. This evolving relationship between technology and fashion has recently captivated the industry, representing a profound shift. The fusion of fashion and technology holds a multifaceted allure, grounded in several compelling factors. Machine learning, a subfield of artificial intelligence, possesses the extraordinary capacity to extract intricate patterns from vast datasets, making it an ideal tool for decoding the complexities of fashion. From predictive analytics that anticipates the next big trend to personalized shopping experiences that cater to individual tastes, the potential applications are manifold.

One of the most captivating developments in the fashion domain in recent times is the emergence of fashion item retrieval systems, especially in the context of composite outfits. As the number of items within each garment category increases, the potential combinations for outfits grow exponentially. Given the typically vast size of fashion inventories, the sheer magnitude of possible outfits that can be curated from these items becomes orders of magnitude greater. The task of mining fashion ensembles from an extensive inventory poses significant challenges, underscoring the necessity for intelligent fashion recommendation techniques [1]. Furthermore, the concept of employing prompts for the purpose of suggesting fashion apparel is relatively new in this field, particularly in the context of recommending multiple harmonious items simultaneously. Consequently, our objective was to address this challenge.

AIABI 2023: 3rd Italian Workshop on Artificial Intelligence and Applications for Business and Industries, November 9, 2023, Milano, Italy

*Corresponding author.

dungnqhe160727@fpt.edu.vn (Q. Nguyen); namphhe160714@fpt.edu.vn (H. Pham); hungddhe160670@fpt.edu.vn (D. Dao); manhdqhe153129@fpt.edu.vn (Q. Do); hatv14@fe.edu.vn (V. Tran)

1.1. Content-based Fashion Retrieval

Content-based fashion image Retrieval (CBFIR) methods retrieved the desired fashion items or products from the queried reference in the form of image, text, or visual clue. The predominant focus within this task revolves around the utilization of referenced images or multimodalities (i.e., image and text) to retrieve desired fashion products for a user. Rubio et al. leverage both the images and textual metadata and propose a joint multi-modal embedding that maps both the text and images into a common latent space, helping effectively perform retrieval in this space [2]. Shin et al. propose a style feature extraction (SFE) layer that decomposes the clothes vector into style and category [3]. They append the layer to the Siamese CNN and train with a loss function composed of softmax loss, contrastive loss, and center loss to predict stylish matching clothes effectively. In recent times, contrastive learning has emerged as a prominent method for acquiring meaningful representations of concepts within the field of machine learning. This approach is grounded in the notion that concepts with semantic connections (for instance, two images of the same object captured from different angles) should exhibit similar representations, whereas unrelated concepts should be distinctly represented. Moreover, CLIP was introduced which represents a multimodal neural network for vision and language, trained using contrastive learning to establish associations between visual concepts and text [4]. Specific to the fashion industry, Chia et al. trained their CLIP model on a fashion dataset containing 800K products [5]. The model, called FashionCLIP, is shown to learn general concepts to be transferable across tasks in the domain. We leverage this model to retrieve fashion items from a textual description.

1.2. Outfit Recommendation using Hash learning

Recent years have seen growing interest in developing intelligent fashion recommendation systems to help users discover and purchase clothing and accessories that match their personal style. The number of possible outfits grows exponentially with the number of items in each garment category. Two ways of evaluating the compatibility of outfit items have been proposed. One approach is to use the pairwise model compatibilities between fashion items, e.g., Siamese network [6], functional factorization [7]. The other one seeks to model high-order relations among the items of an outfit, e.g., a recurrent neural network [8].

Hashing techniques that learn data-driven binary codes have become popular for enabling efficient similarity search in large-scale multimedia retrieval tasks. The aim is to maintain the nearest neighbor relation of the original space in the hamming space. The basic idea is to preserve the similarity, i.e., to minimize the gap between the similarity computed in hash-coded space and the similarity in the original space. Many methods have been introduced by learning real-valued embedding and then taking the sign of the values to obtain binary codes.

Due to the huge amount of fashion items, efficiency becomes an extremely important problem within a practical recommendation system. Learning to hash has been extensively studied for efficient image retrieval. This network models outfit compatibility through pairwise interactions and employs the weighted hashing technique [9] for matching users and items.

2. The Proposed Approach

Fig. 1 illustrates the architectural framework of FashionCLIP, which can be delineated into two distinct phases. In the initial phase, the image encoder undertakes the task of mapping all the garment images contained within the database into a vector space characterized by a dimensionality of 1024. Subsequently, these resulting vectors are persistently stored within the database. In the second phase, when a user submits a query, the text encoder proceeds to project the query into a vector sharing the same dimensional characteristics as the image embedding vector. The prompt embedding vector is then subjected to a dot product operation with all of the image embedding vectors, thereby facilitating the identification of the most compatible garment. FashionCLIP uses transformers with the architecture

modifications described in [10] as the text encoder. The image encoder is a variant of the Vision Transformer (ViT) model [11].

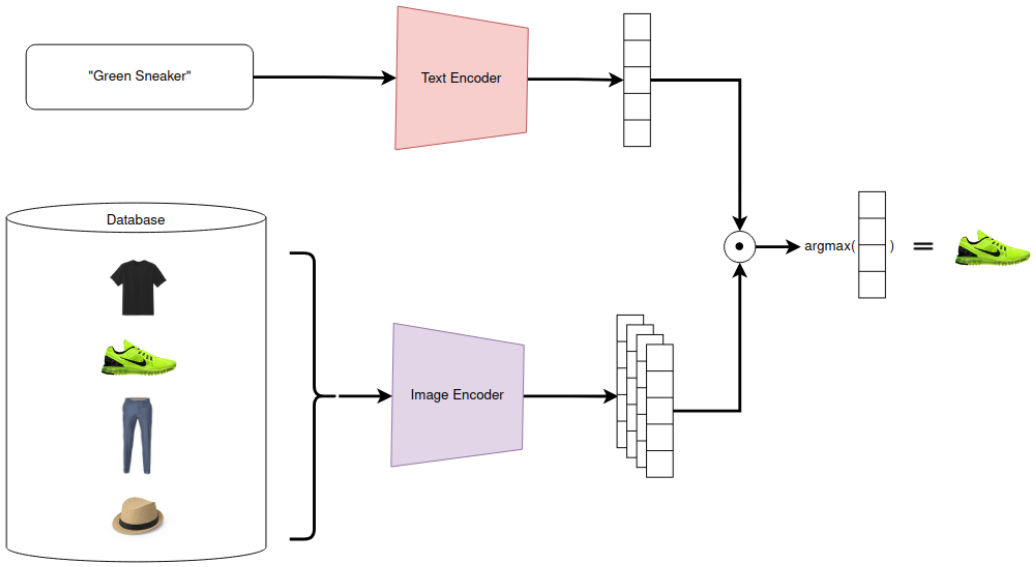

Figure 1: FashionCLIP architecture.

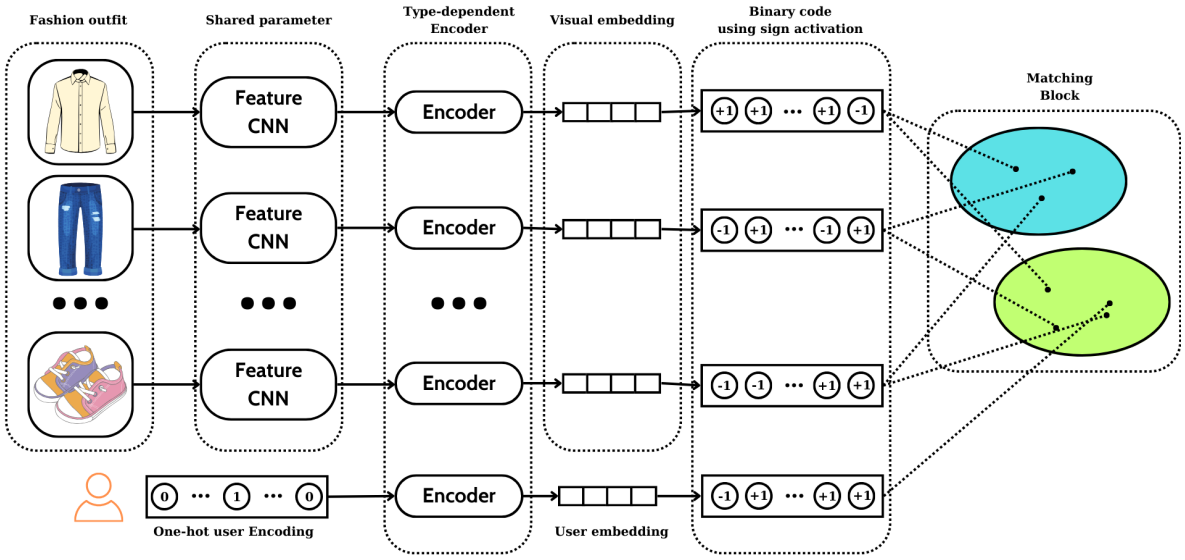

Figure 2: Fashion hashing network architecture.

Also, the architecture of the hashing fashion model is shown in Fig. 2. It includes three components: a feature network that extracts features, multiple type-dependent hashing modules that learn binary codes, and a matching block that predicts preference scores. Each user is represented by a one-hot vector indicating their index. Convolutional networks extract image features. Textual information can optionally be used. Items from different categories and users are treated as different types. The hashing modules contain fully connected layers with a sign function for binarization. The matching block computes the preference score given the binary codes. The final score consists of two terms: one considers item compatibilities and one incorporates users' tastes.

Our pipeline can be delineated as follows: given a textual prompt from a user, we employ FashionCLIP to procure the foremost fashion products that align with the given prompt. These obtained images are treated as a compact database, wherein all the primary apparel items are employed as queries for the

hashing fashion model. When these queries are presented, the hashing model is tasked with retrieving supplementary items from diverse categories such as bottoms, bags, outerwear, and shoes, with the aim of designing a cohesive ensemble.

3. Experiments

3.1. Dataset

The Polyvore dataset provides a large-scale corpus for research on fashion outfit composition. It contains over 1 million user-created outfits compiled from Polyvore, a popular fashion community website. Each outfit includes fashion items of different categories such as tops, bottoms, and shoes that are put together by Polyvore users. The dataset includes rich item metadata, e.g., product images, descriptions, brands, categories, and user engagement statistics. Since its release, Polyvore has facilitated research on outfit compatibility learning and fashion recommendation systems.

3.2. Demonstration

We employ the FastAPI library for the deployment of our hashing network, while the implementation of our model is realized within a web application using the Streamlit library. The model retrieves images from a PostgreSQL database comprising approximately 500 randomly selected images sourced from the Polyvore dataset. We assess the model's performance on a modest computing platform equipped with an Nvidia GeForce GTX 1050Ti GPU, 8GB of RAM.

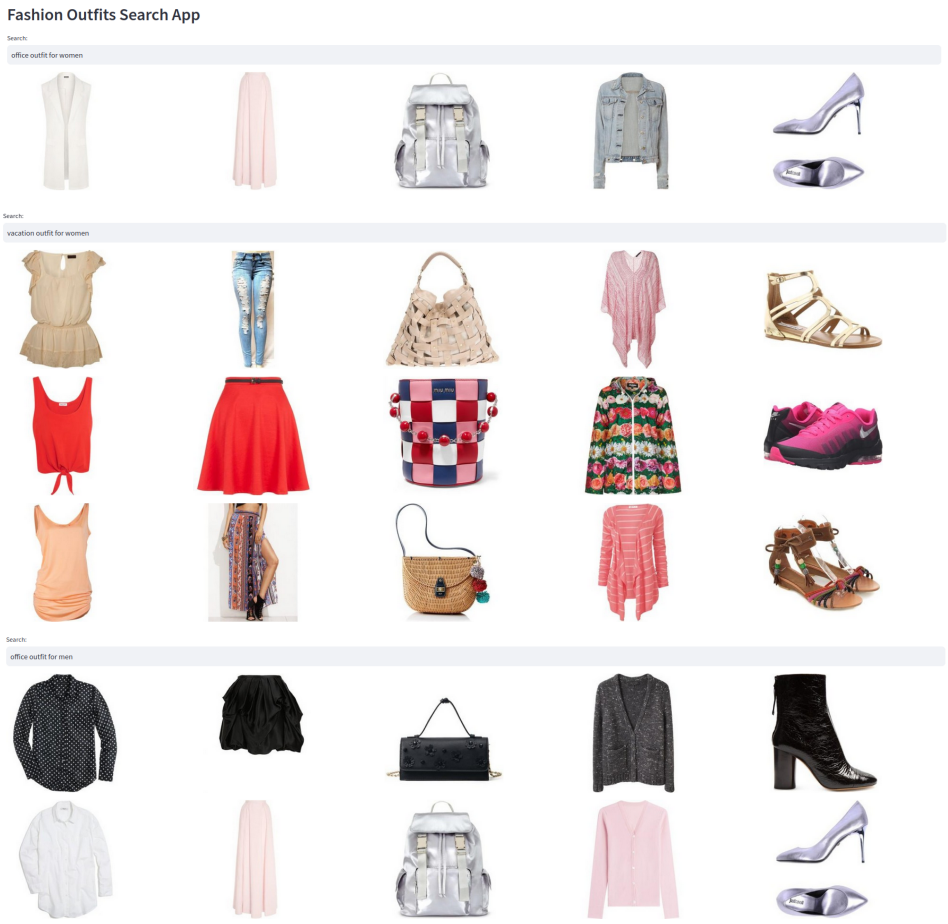

Figure 3: An illustration of the proposed graph partition approach.

Fig. 3 illustrates some prototypical examples of query-based retrieval scenarios, in which a user inputs

a query into the search interface, thereby triggering the system to retrieve ensembles from the database that are compatible with the provided query. Each row of the output exhibits an ensemble comprising five distinct garment categories, namely, top, bottom, bag, outerwear, and shoe. In these instances, the application presents a maximum of three attire recommendations for each given prompt. Notably, the hashing network exhibits superior performance in the context of generating attire corresponding to textual descriptions of female outfits.

Conversely, when tasked with generating recommendations for male outfits, there is an observable tendency for the hashing network to erroneously categorize certain items, particularly within the categories of bottoms and shoes, as female garments, as exemplified in the final query. This discrepancy arises due to the inherent bias within the training dataset, which predominantly consists of female-oriented products within the bottom and shoe categories.

4. Conclusion

In this work, we study how to apply the CLIP model for retrieving the image based on user prompts and study how to utilize the hashing technique for efficient personalized fashion outfit recommendations. Although there are numerous ways to represent the compatibility of outfits, this problem needs to be well handled to fit into hashing optimization. The system performs well in practice, however, it is not an end-to-end solution. Future methods can be proposed for better accuracy or more efficient optimization.

References

[1] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, B. Zeng, Learning binary code for personalized fashion recommendation, in: CVPR, 2019, pp. 10562–10570.

[2] A. Rubio, L. Yu, E. Simo-Serra, F. Moreno-Noguer, Multi-modal joint embedding for fashion product retrieval, in: ICIP, IEEE, 2017, pp. 400–404.

[3] Y.-G. Shin, Y.-J. Yeo, M.-C. Sagong, S.-W. Ji, S.-J. Ko, Deep fashion recommendation system with style feature decomposition, in: ICCE-Berlin, 2019, pp. 301–305.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[5] P. J. Chia, G. Attanasio, F. Bianchi, Contrastive language and vision learning of general fashion concepts, Scientific Reports 12 (2022) 18958.

[6] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, Learning visual clothing style with heterogeneous dyadic co-occurrences, in: ICCV, 2015, pp. 4642–4650.

[7] Y. Hu, X. Yi, L. S. Davis, Collaborative fashion recommendation: A functional tensor factorization approach, in: Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 129–138.

[8] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, Learning type-aware embeddings for fashion compatibility, in: ECCV, 2018, pp. 390–405.

[9] L. Zhang, Y. Zhang, J. Tang, K. Lu, Q. Tian, Binary code ranking with weighted hamming distance, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 1586–1593.

[10] A. Radford, J. Wu, R. Child, D. Luan, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv preprint arXiv:2010.11929 (2020).