Recommendation System for Enhanced Customer Experience: A Novel Image-to-Text Method

Mohamaed Foued A.
The University of Carthage
Higher School of Communications of Tunis, Tunisia
mohamedfoued.ayedi@supcom.tn

Hiba B.
The University of Carthage
Higher School of Communications of Tunis, Tunisia
hiba.bensalem@supcom.tn

Soulaimen H.
Stile.ai
soulaimen.hammami@stile.ai

Ahmed B.
Department of Computer Science
Engineering
Qatar University, Doha, Qatar
abensaid@qu.edu.qa

Rateb J.
Department of Computer Science
Engineering
Qatar University, Doha, Qatar
rateb.jabbar@qu.edu.qa

Abstract

Existing fashion recommendation systems encounter difficulties in using visual data for accurate and personalized recommendations. This research describes an innovative end-to-end pipeline that uses artificial intelligence to provide fine-grained visual interpretation for fashion recommendations. When customers upload images of desired products or outfits, the system automatically generates meaningful descriptions emphasizing stylistic elements. These captions guide retrieval from a global fashion product catalog to offer similar alternatives that fit the visual characteristics of the original image. On a dataset of over 100,000 categorized fashion photos, the pipeline was trained and evaluated. The F1-score for the object detection model was 0.97, exhibiting exact fashion object recognition capabilities optimized for recommendation. This visually-aware system represents a key advancement in customer engagement through personalized fashion recommendations.

1 Introduction

The rapid expansion of E-commerce has transformed online retail, with global sales reaching $4.28 trillion in 2020 (1). This growth is attributed to increased internet access, improved logistics, shifting consumer behaviors, and greater product variety online (2). However, the vast assortment has complicated purchase decisions for fashion shoppers. Inaccurate recommendations lead to dissatisfaction and lower conversion rates (3). Therefore, precise, personalized recommendation systems are critical for fashion E-retailers. Traditional systems relying on metadata, static images, and collaborative filtering struggle to capture nuanced aesthetics (4). Recent research underscores the need for visually-aware, personalized recommendations in fashion (5) (6). Advances in computer vision and deep learning enable richer representations of fashion images and individual preferences. Building on precedents like Google Image Search and Amazon's system (12). This study proposes a novel end-to-end fashion product recommendation system leveraging computer vision and deep learning models. The system incorporates object detection using YOLO-v8 (7) and

5th Deep Learning Indaba Conference (DLI 2023).

product classification through FashionClip (8) to understand customers' visual preferences from product images. Descriptive captions are generated using BLIP (10), an image captioning technique. Retrieval and recommendations are enabled through efficient search over a product catalog scraped from retailers globally using OpenSearch (11).

2 Methodology

In this section, we present the details of our proposed recommendation system. Figure 1 depicts the details of the proposed recommendation system. The process starts with an outfit image as input to a fine-tuned object detection model, enabling us to isolate each distinct product. The cropped images are then passed through a zero-shot classification model to generate the precise label of the product. Moving forward, the product image, along with its label, is passed to an image captioning model to generate detailed descriptions of the item's pattern and color.

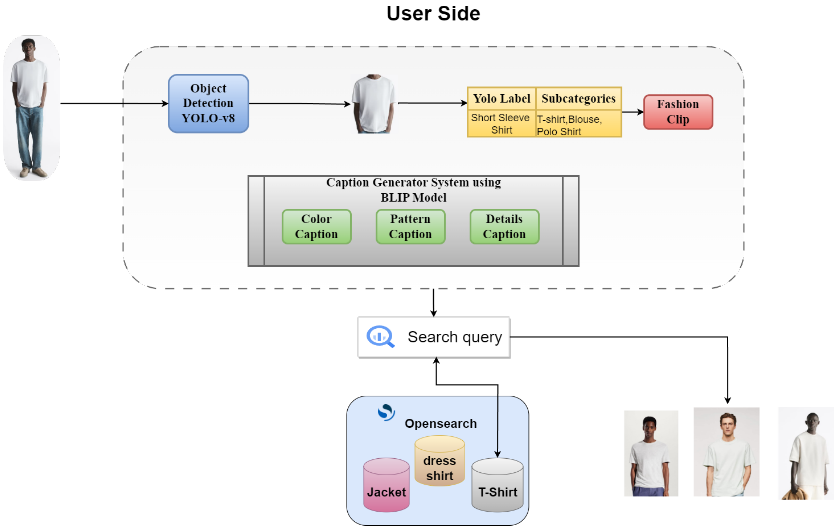

Figure 1: Overview of the proposed architecture.

To manage our data, we employed OpenSearch. In this setup, product labels serve as indexes for our search clusters. The final step in this procedure involves locating similar items. In the following sections, we will describe each part of the system in detail.

2.1 Data preparation

We started with curating and balancing a dataset of 111,824 images sourced from web scraping. It is essential to have a balanced dataset when training object detection models like YOLO-v8 to avoid any class bias.

2.2 Training and evaluation process

The training procedure consisted of fine-tuning YOLO-v8 model for 100 epochs on NVIDIA RTX 4090 for approximately 35 hours. Our dataset was divided, with 90% dedicated to training and the remaining 10% used for testing and validation. We monitored three loss metrics: box loss, class loss, and defocus loss. Box loss relates to errors in bounding box prediction, class loss deals with errors in class prediction, and defocus loss considers errors in focus prediction.

2.3 Object detection

After training, we used our model to detect objects in input images. This allowed us to categorize and crop images of products into distinct items, such as long sleeve tops, short sleeve tops, long sleeve outerwear, trousers, and shorts. The main goal of this step was to analyze each part of an outfit independently.

2.4 Products Classification

For precise classification of fashion products, subcategories were created for each YOLO-recognized class. FashionClip (13), a CLIP model (14) adaptation, was then used. It was trained on a dataset of over 700,000 image-text pairs from Farfetch (9), a well-known luxury fashion retailer, and excels in "zero-shot generalization". Each cropped image, along with its YOLO-v8 label and subcategory, is processed by FashionClip, which generates the most fitting label within the subcategory by comparing image and labels embeddings and selecting the label with the highest cosine similarity score.

2.5 Caption Generation

In this step, we employ cropped images, coupled with their new labels, as input to an image captioning model to generate precise descriptions. We opted for BLIP (15), a bootstrapping Language-Image Pretraining for Unified Vision, for its precise alignment between visual and textual content. BLIP employs the Vision Transformer (ViT) approach (16), to segment the input image into patches and encodes them as a sequence of embeddings. The BLIP components, Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), and Language Modeling Loss (LM) link, evaluate, and generate text descriptions respectively. To slightly augment the output from BLIP, we used a prompt-based approach. At the start of each caption, prompts such as "this label features" are included, directing the generation process toward specific outfit aspects.

2.6 Product Similarity Recommendations

We enhanced our recommendation system by implementing data categorization across OpenSearch clusters. By using the product label from Fashion Clip (FC) and the captions generated using BLIP, we employed the match query for full-text searches. This approach considers word proximity and generates ranked recommendations based on matching scores. The integration of Okapi BM25, a powerful ranking function, within OpenSearch resulted in performance improvements, delivering faster and more precise suggestions (17).

3 Experiments

We rigorously evaluate our fashion product recommendation system through a series of quantitative and qualitative experiments. The experiments are conducted on dedicated test datasets, and we provided category-wise evaluations to gain comprehensive insights into our system's capabilities.

3.1 Object Detection Results (YOLO-v8)

As illustrated in Figure 2, the YOLO-v8 model achieved impressive results in object detection, with an average of 0.97 accuracy across five categories. Figure 3 further exhibits the Precision-Recall Curve for our model. However, occasional challenges were encountered when differentiating between long sleeve tops and outerwear.

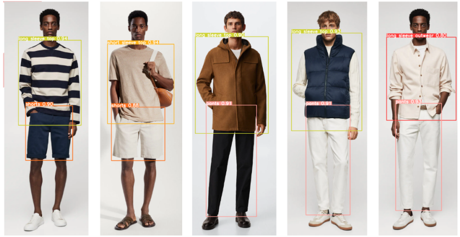

Figure 2: Object Detection Results with YOLO-v8 on Men Outfits.

3.2 End-to-End Evaluation: image containing examples of generated product similarities

BLIP, the caption generation model, excelled in providing intricate details about fashion products, including colors, patterns, and designs. This capability enhanced the quality of our products rec-

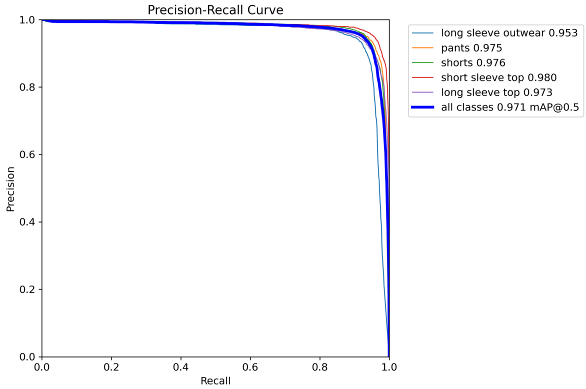

Figure 3: Precision-Recall Curve for Object Detection

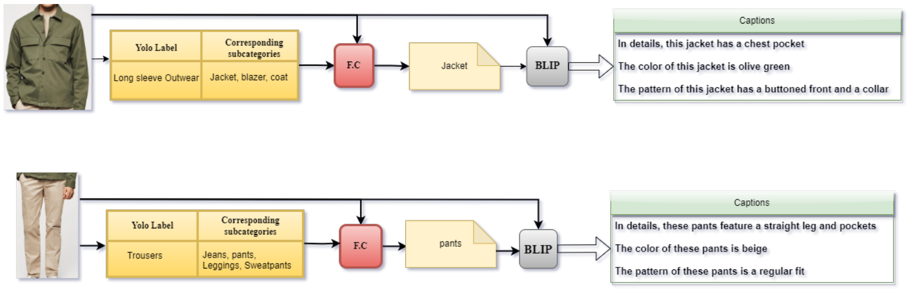
Yolo Label
Corresponding subcategories
Long sleeve Outwear
Jacket, blazer, coat
F.C
Jacket
BLIP
Captions
In details, this jacket has a chest pocket
The color of this jacket is olive green
The pattern of this jacket has a buttoned front and a collar
Yolo Label
Corresponding subcategories
Trousers
Jeans, pants, Leggings, Sweatpants
F.C
pants
BLIP
Captions
In details, these pants feature a straight leg and pockets
The color of these pants is beige
The pattern of these pants is a regular fit

Figure 4: Captions generation results

ommendations. Additionally, the fashion clip's accurate classification labels were instrumental in creating effective search clusters. Figure 4 shows an example of FashionClip and BLIP results.

Overall, our system consistently delivers strong results for fashion products, and Figure 5 exemplifies some similarity results, affirming the system's effectiveness in providing relevant fashion product recommendations. For each query image, we present the top retrieved results.

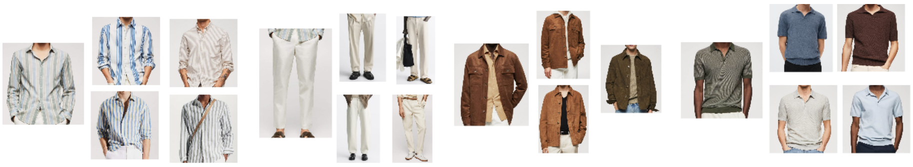

Figure 5: End-to-end results of the visual search system.

4 Conclusion

Our research has successfully engineered a novel fashion product recommendation system, with promising results in terms of accuracy. This system, utilizing a fine-tuned process of object detection, zero-shot classification, and image captioning, has the potential to transform online fashion retail. Nevertheless, we have encountered certain challenges during this process. The success of our system depends on the quality of the input image and the accuracy of each model component. If any component fails to identify or describe an item correctly, it could have a significant impact on the results. Going forward, our future work holds great potential. We aim to expand our system's capabilities to suggest not only similar items but also complementary outfit items. This would enrich the user shopping experience by providing comprehensive outfit suggestions.

References

[1] Company, M.R. (2021). Global ecommerce update 2021. Retrieved from `https://www.emarketer.com/content/global-ecommerce-update-2021`

[2] Smith, R. (2019). Ecommerce product return rate – statistics and trends. Retrieved from `https://www.invespcro.com/blog/ecommerce-product-return-rate-statistics/`

[3] Park, E., Han, X., Berg, T.L., & Berg, A.C. (2017). Combining multiple sources of knowledge in deep CNNs for action recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1222-1231). IEEE.

[4] Gajic, B., & Balzano, L. (2018). Fashion imagery feature extraction and style recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 0-0).

[5] Jagadeesh, V., Piramuthu, R., Bhardwaj, A., Di, W., & Sundaresan, N. (2014). Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1925-1934).

[6] Corbière, C., Bengio, Y., Lecun, Y., & Pal, C.J. (2017). Leveraging weakly annotated data for fashion image retrieval and label prediction. *arXiv preprint arXiv:1712.02201.*

[7] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767.*

[8] Gao, T., Jiang, X., Zhang, T., Yao, Y., Tay, Y., Li, S., ... & Sun, H. (2021). Fashionclip: Style-aware vision-language modeling for fashion. *arXiv preprint arXiv:2204.03972.*

[9] Fashion Product Images Dataset Retrieved from `https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset`

[10] Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Wang, L., ... & Batra, D. (2021). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4929-4938).

[11] Gormley, C., & Tong, Z. (2015). Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. "O'Reilly Media, Inc.".

[12] Tashjian, R., (2019) "How Jennifer Lopez's Versace Dress Created Google Images," *GQ*, Retrieved from `https://www.gq.com/story/jennifer-lopez-versace-google-images`

[13] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., (2021) "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020.*

[14] Chia, P. J., Attanasio, G., & Tagliabue, J., (2022) "Contrastive language and vision learning of general fashion concepts," *arXiv preprint arXiv:2204.03972.*

[15] Li, J., Li, D., Xiong, C., Hoi, S., (2022) "Salesforce Research BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *arXiv preprint arXiv:2201.12086.*

[16] Opster Team, "Elasticsearch Match, Multi-Match, and Match Phrase Queries," Retrieved from `https://opster.com/guides/elasticsearch/search-apis/elasticsearch-match-multi-match-and-match-phrase-queries/#Match-phrase-query`.

[17] Amati, G., (2009) "BM25," In: Liu, L., Özsu, M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA.