Fashion Product Recommendation System. "StyleSelect: A Novel Approach for Personalized Fashion Product Recommendations"

Sejal Kanoje
C-03
GH Raisoni College of Engineering
Nagpur

Tanvy Parate
C-13
GH Raisoni College of Engineering
Nagpur

Thorvi Parchand
C-14
GH Raisoni College of Engineering
Nagpur

Abstract – The study suggests a brand-new computer vision-based method for fashion product recommendations called StyleSelect. It tackles the difficult task of making recommendations for related fashion goods based on a user's query and the Product Display Page (PDP) that corresponds to the query. [1][2].A triplet network-based picture embedding model, human keypoint identification, pose classification, article localization, and object detection are some of the components that make up the suggested method. [1].In addition to the main article that corresponds to the query, the algorithm seeks to suggest related fashion items matching the full collection of fashion goods worn by a model in the PDP full-shot image. This strategy enhances customer experience and engagement while also encouraging cross-sells to increase income. [1][2]. Myntra, an established online fashion e-commerce company, assessed the approach and found that it performed well in terms of suggesting related products for both the main query and user-generated content. [1] [2].

I. Introduction

The research paper titled " Fashion Product Recommendation System " presents a novel approach named 'StyleSelect' for proposing related fashion items within the framework of customized fashion [1][2]**[6]**. StyleSelect leverages computer vision techniques, including human keypoint detection, pose classification, article localization, object detection, and a Triplet network-based image embedding model **[1]**. The goal of StyleSelect is to offer personalized suggestions for similar fashion products based on user queries and model images, providing a holistic shopping experience **[1]**.The approach not only identifies similar articles to the primary query but also extends to recommend akin products for all items adorned by the model, enhancing cross-selling opportunities and customer satisfaction **[4]**. The methodology seamlessly extends to User Generated Content (UGC), enabling recommendations for fashion items uploaded by users **[2]**.The research paper evaluates the effectiveness of StyleSelect on Myntra, a leading online fashion retailer **[1]**. The proposed methodology ensures accurate detection, localization, and retrieval of fashion articles, facilitating effective recommendation of similar fashion products to users based on their queries **[2]**. The research paper highlights the growing demand for personalized fashion recommendations in the context of social media influencers and e-commerce platforms **[2]**.

II. Background and Related Work

The research paper addresses the increasing demand for personalized fashion recommendations in the era of social media influencers and e-commerce platforms [1]

[2]. The proposed approach, StyleSelect, leverages computer vision techniques, including object detection, embedding learning, and human keypoint estimation [1] [2].

1. Object detection is performed using the Mask RCNN model to identify and localize fashion articles [2].
2. Embedding learning is employed using a Triplet network-based model to represent fashion articles in a common embedding space [2].
3. Human keypoint estimation is utilized to identify key-points in full-shot images, enabling accurate detection and retrieval of fashion articles [1][2].
4. The research paper focuses on the fashion application of StyleSelect, catering to users' desires for curated lists of fashion items akin to those worn by models featured on digital platforms [1][2].

III. Proposed Methodology

The proposed methodology, named 'StyleSelect', aims to recommend similar fashion products based on user queries and corresponding Product Display Pages (PDP) **[1]**.

The methodology consists of the following components:

1.**Human keypoint detection** is performed to identify key-points in the full-shot look image **[2]**. By confirming that the head and ankle keypoints are present in a set of product display photographs, it is utilized to identify the full-shot look image [2].It has been demonstrated that the method of Xiao et al. works better for human keypoint estimate than other competing techniques [3]. For feature extraction, the technique uses a ResNet backbone. Next, it uses deconvolutional layers and a 1x1 convolutional layer to provide predicted heatmaps for keypoint detection [4].

2.Pose classification categorizes the image into front, back, left, right, or detailed shot **[2]**.This pose classification step is crucial in accurately identifying and recommending analogous fashion products corresponding to the entire ensemble depicted in model images [1].The pose classification is performed to categorize the image into different angles and perspectives, which helps in accurately detecting and localizing fashion articles worn by the model [1].

3.Article localization and object detection are conducted using the Mask RCNN model to detect and localize fashion articles **[3]**.Article localization and object detection in the research paper is performed using the **Mask RCNN model**. A tailored training dataset including approximately 7-9k training photos and 800 test images for each category of fashion article is used to train the model. By giving the bounding box position and classification for about 20 different types of clothing, it can identify and locate fashion goods. All classes saw an average mAP of 78 for the model, with some top wear classes seeing values as high as 92. Furthermore, an active learning environment is included, whereby internal taggers find samples that were incorrectly classified and utilize them to retrain the model, resulting in additional gains in performance measures**[1][2].**

4.A triplet network-based image embedding model - Fashion articles are represented in a shared embedding space using an image embedding model based on triplet networks [3].In the study paper, a triplet network-based image embedding model is used to represent fashion articles in a shared embedding space. Three identical Convolutional Neural Networks (CNN) with common weights make up the triplet network. Moving away from different images and toward the embeddings of semantically comparable images is the goal of training the network. (xa, xp, xn) represents the embeddings for a triplet of pictures, which consists of an anchor image, a positive image, and a negative image. The embeddings are utilized to compute picture similarity in order to get related products from the catalog database. The model is trained on a large-scale collection of real-world fashion products. This method makes it possible to identify dissimilar articles in the embedding space and group similar articles together.[1]

Active learning feedback is incorporated to identify misclassified examples and improve model performance through re-training **[3]**.

On Myntra, the suggested method's effectiveness is assessed. A leading online fashion retailer, to cater to users' desires for curated lists of fashion items **[4] [5]**. The methodology seamlessly extends to User Generated Content (UGC), enabling recommendations for fashion items uploaded by users **[4]**. The proposed method builds upon the "Buy Me That Look" approach and leverages deep learning techniques for fashion article detection and retrieval. It consists of three main stages:

1. Full-shot image detection with front facing:

Human Keypoint Detection: Deep learning models are employed to identify keypoints (e.g., elbows, wrists, ankles) on the human body in the image. Reference: Xiao et al. (2018) [3] or similar keypoint detection techniques.
Pose Classification: Based on the detected keypoints, the model classifies the image pose as front-facing, back, left, right, or a detailed close-up shot. This filtering ensures the system focuses on full-body images for optimal clothing detection accuracy.

1. Fashion Article Detection and Localization:

Image Pre-processing: The input image undergoes pre-processing steps like resizing and normalization to ensure compatibility with the object detection model.
Mask R-CNN for Detection and Localization: The pre-processed image is fed into a Mask R-CNN model [5] trained to detect individual fashion articles (shirts, pants, etc.) and generate segmentation masks around them. This not only identifies the clothing items but also precisely locates them within the image.

2. Image Retrieval:
Once trained, the network generates an embedding for each extracted fashion article in the query image.
By comparing the embedding of the query article with the embeddings of articles in a database (e.g., online store's product catalog), the system retrieves visually similar clothing items.

Creating Embeddings for Different Article Types:

Extracted Fashion Articles: The segmentation masks from Mask R-CNN correspond to individual clothing items.

3. Triplet Network Architecture: A triplet network is employed to learn meaningful representations (embeddings) for the fashion articles. This network takes three images as input:
An anchor image (a clothing item from the query image)
A positive image (an image containing a similar clothing item)
A negative image (an image containing a dissimilar clothing item) With respect to the embedding space, the network's goal is to reduce the distance between the embeddings of the anchor image and the positive picture, and to maximize the distance between the embeddings of the anchor image and the negative image.
This training process compels the network to learn effective representations that capture the essence of each clothing item. [6] for FaceNet or similar triplet network architectures can be explored.

Triplet Loss Minimization and Embedding Loss Normalization: Techniques like triplet loss and embedding loss normalization are utilized during training to ensure the triplet network effectively learns discriminative embeddings [3] [9].

IV. Implementation

This section explores the characteristics of the suggested fashion article detection and retrieval method's implementation.

Model Implementations:

Human Keypoint Detection: Existing pre-trained models like OpenPose or AlphaPose can be employed for this task. These models typically take an image as input and output the locations of key human body points (e.g., joints)
Mask R-CNN: Using a Mask R-CNN model that has already been trained, such as Mask R-CNN with ResNet-50 backbone [5].
This model can be fine-tuned on a fashion article detection dataset to improve performance on clothing items.

Triplet Network: Implement a triplet network architecture. Libraries like TensorFlow or PyTorch offer modules for building custom network architectures.

Libraries and Frameworks:

1. **TensorFlow or PyTorch:** These popular deep learning frameworks provide the foundation for building and training the neural network models.
2. **Keras (optional):** Keras can be used as a high-level API on top of TensorFlow for a more user-friendly development experience. (Ref: [2])

Training Process:

1. **Model Training:**

To fine-tune the Mask R-CNN model for fashion article detection, train it using the provided dataset. Train the triplet network using the generated triplets. The training process involves minimizing the triplet loss and embedding loss functions.

2. **Data Preparation:**

Prepare a dataset of images containing labeled clothing items. (e.g., DeepFashion [8])
Pre-process the images by resizing and normalizing them for compatibility with the models.
Extract triplets of images (anchor, positive, negative) for each clothing item in the dataset. The positive image should contain a similar garment, while the negative image should contain a dissimilar one.

3. **Inference:**

1. **Pre-process the query image.**
2. **Run human pose estimation to identify the pose.**

4. **Image Retrieval:**

Compare the embedding of each extracted clothing item with the embeddings in the database (e.g., product catalog).
Retrieve items with the most similar embeddings based on a predefined threshold.
If the pose is front-facing, proceed:
Locate and identify clothing articles in the image using the Mask R-CNN model.
For each detected item:
Extract the corresponding region of interest (ROI) from the image.
Generate an embedding for the ROI using the trained triplet network.

V. Future Scope

Integration of additional features like occasion-specific recommendations and finer attributes to enhance the model's contextual relevance [2].
Refinement of recommendation accuracy through advanced personalization techniques such as collaborative filtering and natural language processing [1].Enriching user experiences by incorporating multi-modal data sources and interactive features like virtual try-ons [6]. Establishing a robust framework for continuous model evaluation to ensure responsiveness to evolving user preferences [2]. Incorporating article attributes and occasion-based filtering can enhance the product search and recommendation process.**[1]** Further improvements can be made in the active learning component to enhance the performance of the fashion article detection and localization.**[2]** Pose categorization, object detection, and human keypoint detection can all be made more accurate and efficient by investigating different computer vision models and approaches.[3] Recommendations for fashion products can be made more personalized by incorporating user preferences and feedback.[1] Performing user research and obtaining input to assess the suggested method's efficacy and user satisfaction in practical settings.[4]

Conclusion-

In the study paper, a novel computer vision-based method named ShopLook is proposed for fashion product recommendation based on the matching Product Display Page (PDP) image and a user inquiry. The technique uses a triplet network-based picture embedding model, active learning feedback, object detection, pose categorization, and human keypoint recognition.The suggested approach has been tested at Myntra, a well-known online fashion marketplace, and the results have been encouraging in terms of suggesting related fashion pieces for the full assortment of clothing that a model is wearing.

Following research may concentrate on integrating article characteristics with occasion-based filtering to strengthen and optimize the process of product search and recommendation. Additionally, exploring user feedback and preferences, as well as conducting user studies, can further enhance the personalization and effectiveness of the fashion product recommendation system.

References-

1. arXiv:2008.11638v2 [cs.CV] 6 Apr 2021 https://dl.acm.org/doi/abs/10.1145/3394486.3403372
2. Praveen P., Rama B(2020). "An Optimized Clustering Method To Create Clusters Efficiently" 2020 Journal Of Mechanics Of Continua And Mathematical Sciences , ISSN (Online) : 2454 -7190 15(1), pp 339-348
3. Xiao, B., Wu, H., Wei, Y., et al. (2018). Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany. Retrieved from: https://arxiv.org/abs/1804.06208
4. IOP Conf. Series: Materials Science and Engineering 981 (2020) 022073 IOP Publishing doi:10.1088/1757-899X/981/2/022073
5. He, K., Zhang, X., Ren, S., et al. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy.Retrieved from:https://arxiv.org/abs/1703.06870
6. Schroff, F., Kalenichenko, D., Philbin, J. (2020). FaceNet: A Unified Embedding for FaceRecognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. Retrieved from: https://arxiv.org/abs/1503.03832
7. Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE transactions on pattern analysis and machine intelligence (TPAMI), 41(4):871–885, 2020
8. Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In Proceedings of the 3rd ACM conference on International conference on multimedia retrieval (ICMR), pages 105– 112, 2013.
9. Simonyan, K., Zisserman, A. (2021). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556. Retrieved from: https://arxiv.org/abs/1409.1556
10. Szegedy, C., Liu, W., Jia, Y., et al. (2017). Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. Retrieved from https://arxiv.org/abs/1409.4842