A Multimodal AI System for Personalized Fashion Recommendation

Carlota Fernández del Riego
Universidad Intercontinental de la Empresa
A Coruña
carlota.fernandez.01@uie.edu

***Abstract*—Fashion is an ever-evolving field where inspiration and personalization are key factors for user interaction. However, current digital tools often lack dynamic and personalized recommendations that integrate visual understanding with natural language interaction. This paper presents a novel AI-based system that combines deep learning in computer vision and natural language processing (NLP) to improve fashion recommendations. Users can interact with a chatbot via conversational queries and images to discover garments suited to their preferences. The system estimates similarity based on visual and semantic features and provides direct links to e-commerce platforms. This approach enhances the relevance of fashion recommendations and bridges the gap between user intent and product discovery.**

***Index Terms*—Fashion Recommendation, NLP, Computer Vision, Deep Learning, Chatbots**

I. Introduction

Fashion is a cultural and social occurrence that is constantly evolving and is influenced by trends, historical contexts, and personal preferences. Today, with the expansion of social media and e-commerce, people are increasingly looking for inspiration and advice to define their style and select their clothing depending on the occasion they need to attend.

However, most existing platforms lack customised, visually interactive, and easy-to-understand systems to effectively assist the user in this suggestion and choice process. It is for this reason that many people decide to leave aside the technological part and hire people specialized in the fashion sector, but what they do not know is that without technology they would not be able to carry out their work correctly.

The process of choosing a clothing set involves multiple factors, like personal aesthetics, occasion of use, color combination, current trends, and the availability of products in the market. However, there are several solutions based on conventional search engines or categorical filters, have significant limitations, as they lack tools capable of interpreting both natural language and images of all types of garments.

Nowadays there are online catalogues and visual search engines, but these are often limited in terms of customization, semantic understanding, and the ability to suggest relevant alternatives. Therefore, if they do not have face-to-face workers to help each user, the work they sell is not fully advised, and they do not take full advantage of the visual analysis capabilities provided by modern artificial intelligence.

One of the most relevant challenges in fashion recommendation lies in the so-called ”semantic gap” [1]: the discrepancy between the textual description of a garment and its visual representation. Overcoming this problem is essential to offering truly accurate and satisfactory recommendations based on the tastes of each user and all the factors that lie in that outfit.

In this context, the work carried out proposes the development of an intelligent chatbot that integrates natural language processing (NLP), image analysis to recommend similar outfits and garments in an automated way (Computer Vision and Advanced Machine Learning). In addition to suggesting personalized combinations, the system calculates the similarity between different garments based on visual and semantic details and provides direct links to websites where users can purchase the suggested products knowing in advance all their characteristics, including their price (Intelligent Systems).

The main contributions that have been developed for this project are:

- The integration of NLP techniques such as word processing and the use of transformers for complete user understanding.
- The visual interpretation of the complete dataset together with its key characteristics for the recommendation of all types of garments.
- The creation of a search system for similar garments based on images and descriptions based on a similarity graph.
- The design of an interactive experience that connects the conversation with a chatbot, personalized recommendations and the link with online shopping platforms to be able to purchase these garments on their website.

II. Ease of Use

A. Fashion Recommendation Systems

The mixing of artificial intelligence with fashion has grown significantly as an area of research, and this has led to improvements in the field of commerce. The integration of computer vision and language processing has been key to the work of recommendation systems based mainly on image analysis or suggestions based on user history.

In the area of computer vision, techniques such as convolutional networks (CNNs) have been widely used to extract visual features from garments and calculate similarities between them. On the other hand, models such as FashionNet and DeepFashion [2] have enabled significant advances in the recognition and classification of clothing from images, establishing standardized databases that have driven the development of new algorithms. Even so, the existing datasets do not contain the complete characteristics necessary to work with all the information in this field.

In parallel, natural language processing has been explored to improve the interaction between users and recommendation systems to achieve a complete proposal. Recent studies have used transformer-based models, such as BERT, to interpret garment descriptions and user queries, thus improving the relevance of suggestions [3]. Today, these models are the best option when it comes to working with descriptions, allowing the processing of the sentence word by word and consequently solving the problem of processing time.

B. Multimodal Recommendation and Chatbots

Multimodal recommendation systems leverage and integrate multiple types of data to predict and suggest items that fit your preferences [4]. In fact, thanks to combining textual and visual information, they have shown promising results. However, many of these systems are designed for passive search contexts, where the user enters keywords or selects predefined filters instead of interacting in a conversational way. They are simply used to search in a large context for keywords to gather information, but not for the purpose of storing it.

Additionally, the use of chatbots in the commercial field has gained popularity, especially in the online sales sector. However, most chatbots today are limited to basic customer service functions and lack advanced image interpretation capabilities or personalized recommendations based on visual content. In fact, most of these do not have the ability to understand complete sentences, and therefore these systems show different options for the user to select one and send it without the need to write any type of sentence.

C. Gaps and Opportunities

Despite these advances, there is still a large gap in the development of systems capable of understanding natural language in a conversational way, analyzing fashion images accurately, and providing personalized recommendations that integrate both types of information. This lack motivates the development of this project, which seeks to combine advanced natural language processing and computer vision techniques in an intelligent chatbot that improves the fashion research and shopping experience for users.

III. Experimentation

A. Dataset and Preprocessing

For this project, a proprietary dataset has been developed due to the lack of a set of images of clear clothing with all the necessary characteristics to work with it without presenting problems. The dataset created is specific and tailored to the goals of the personalized fashion recommendation. In fact, it is made up of approximately 300 to 500 images of garments, collected manually from Zara's web platform, combining manual processes and web scraping techniques to extract both visual files and product descriptions.

The images have been organized into eleven main categories: heels, accessories, boots, t-shirts, shirts, sweatshirts, sweaters, pants, sneakers, jackets and coats. Each image was stored in PNG format to preserve visual quality and facilitate subsequent stages of image processing. In addition, associated CSV files were created, in which the corresponding tags and extracted textual descriptions were recorded, thus allowing the linking between visual and semantic data.

The data processing performed for this project encompassed both visual and textual elements.

For image processing, processes of reading, RGB color conversion, pixel normalization and resizing to a standard pixel size were applied. These images were used for training two convolutional transfer learning networks: VGG16 [5] and ResNet50 [6], both pretrained on ImageNet. The output of the feature layer of VGG16 was used to generate vector representations of the images, which served for clustering using KMeans and for the construction of a similarity graph, where the most similar garments were connected.

As for the textual data, they were obtained from fashion magazines such as Vogue or Elle and from online articles for which it was necessary to apply web scrapping. The spaCy library was used for tokenization, stopwords removal, lemmatization and normalization of extracted descriptions. From these preprocessed texts, TF-IDF was applied for the extraction of relevant keywords in the fashion domain, and models based on multilingual BERT were used to analyze semantic trends within the textual corpus. Due to the low number of fashion magazines with a high reputation in English, it was necessary to download different copies and translate them so that the analysis of the words could be possible.

In addition, an automatic generation system of outfit descriptions based on generative models such as GPT-2 was implemented, both from processed images and explicit preferences provided by users.

Finally, to build a feedback-based recommendation system, a synthetic set of user ratings on fashion items was generated, and a matrix factorization model was trained using the SVD (Singular Value Decomposition) technique [7], approved with a cross-validation methodology.All the processed dataset and their respective characteristics were stored in CSV and NPY format for later integration into the conversational recommendation engine.

B. System Architecture

The overall architecture of the system was designed with the goal of delivering a personalized fashion recommendation experience through a seamless and easy-to-understand conversational interaction. The system integrates natural language processing, computer vision, and recommendation mechanisms into a single operational flow, applying both intelligent systems and advanced machine learning.

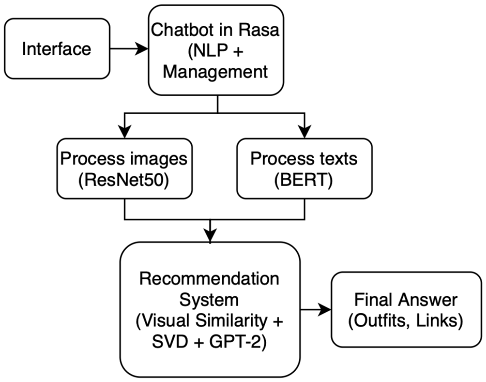
Interface
Chatbot in Rasa (NLP + Management)
Process images (ResNet50)
Process texts (BERT)
Recommendation System (Visual Similarity + SVD + GPT-2)
Final Answer (Outfits, Links)

Fig. 1. General architecture of the proposed system.

For the implementation of the conversational component, the Rasa platform, which is an open-source and asynchronous artificial intelligence framework, was used for the creation of contextual assistants as chatbot applications [8]. This framework was trained to interpret user intent, recognize relevant entities, and manage dynamic dialogues.

Rasa is mainly composed of two fundamental modules: Rasa NLU (Natural Language Understanding) and Rasa Core [9].

1) **Rasa NLU** is responsible for processing textual inputs from users, identifying the *intents* and extracting relevant *entities*. For this project, NLU training was supported by a BERT-based model, optimizing the semantic understanding of fashion queries in natural language.
2) **Rasa Core** manages conversation and dialogue logic through a set of *stories* and *rules*. Stories represent example conversational flows that the assistant must learn to handle, while rules define specific responses to particular conditions.

The configuration of the chatbot was structured in the following components:

- **Domain**: Defines the intents, entities, actions, and responses available in the wizard.
- **NLU**: Contains training examples for intent detection and entity extraction.
- **Stories**: Describes possible conversational paths based on sequences of intents and actions.
- **Rules**: Establishes deterministic actions in specific situations, such as the confirmation of garment selection.
- **Actions**: Define custom actions, such as generating outfit descriptions.

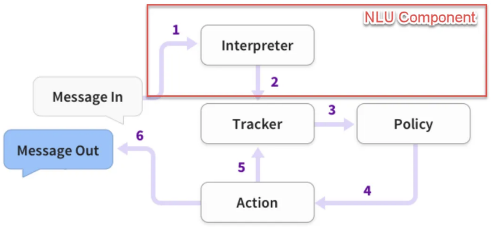

Fig. 2. Rasa chatbot architecture with the NLU part checked.

During the interaction, when the user sends a text message, it is processed through the NLU pipeline, identifying the intent and entities. The system, based on this understanding, provides outfit recommendations, style suggestions and allows access to the other two sections of recommendation by images and similarity graphs to have information on specific garments.

C. Visual Recommendation of Garments

In addition to conversational interaction, the system includes a recommendation module based on the upload of images of garments. When the user provides an image, it is processed by a ResNet50 convolutional network, pretrained on ImageNet, to extract deep visual features. This convolutional lattice has been chosen because it allows very deep lattices (of more than 100 layers) to be trained, successfully controlling the problem of ’gradient fading’ [10]. These visual representations are compared using a previously constructed similarity graph, allowing to find items similar to the garment uploaded by the user. The graph was constructed from features extracted from a set of garments organized by clustering techniques (KMeans), thus optimizing navigation by visual relationships. In addition, the system offers an automatic generation of outfit descriptions based on generative language models such as GPT-2, adapting the suggestions to the declared style, color and occasion preferences.

The multimodal recommendation engine is complemented by an SVD matrix factorization model, which allows garments to be suggested based on simulated user ratings, thus incorporating a collaborative approach that enriches the recommendations. This model is based on a generalization of the eigenvalue decomposition of a matrix, and can be applied to any rectangular matrix, not just square matrices. It is mainly used to find underlying patterns in large texts [11].

This integration of components allows the system to offer a personalized experience based on visual inputs, to finally propose complete outfits, similar garments or alternative styles, improving the fashion discovery experience in an integral way.

D. Garment Similarity Graph

To improve the quality of the visual recommendations and allow an interactive exploration of similar garments, a similarity graph was designed and constructed based on the visual characteristics extracted from the images of garments.

The graph construction process began by extracting visual embeddings using a pre-trained ResNet50 convolutional network, applied to the normalized images of the dataset. From these embeddings, similarities between pairs of garments were calculated using the cosine similarity metric. This metric is based on the measurement of the angle between two documents in the metric space of multiple dimensions [12].

The similarity graph was constructed following the following steps:

- Each node represents a garment in the dataset, tagged with its category.
- For each node, the most similar k garments were connected, where k is a configurable parameter, which establishes the connections by edges along the degree of similarity.

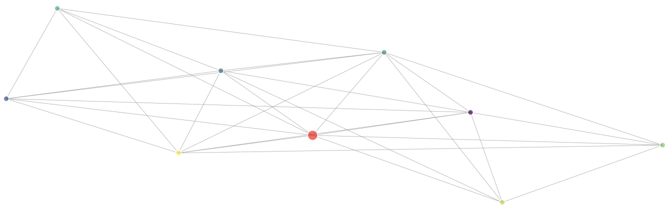

Fig. 3. Similarity graph with highlighted shirt category node.

The construction of the graph was carried out using the NetworkX library, and its persistent storage was implemented using serialization using Pickle, allowing its subsequent loading and use without the need to recalculate the similarities. The primary use of this is to convert a Python object into a byte stream, which can be stored in a file or transferred over a network [13]. To recommend similar garments from an entry-level garment, a Breast-First Search (BFS) wide search algorithm was implemented in the graph. This algorithm allows us to explore neighboring nodes to a certain depth, thus recovering a set of garments closely related to the initial garment of the graph.

In addition, an interactive visualization of the graph was developed using Plotly, allowing the user to explore local subgraphs focused on a specific garment. This way the user can know visually. The visualization includes:

- Representation of nodes (garments) and edges (similarities).
- Direct relationship between the different garments.
- Differentiated colors for each garment category.
- Prominent color when searching for a certain garment.
- Images and features dynamically displayed when selecting a particular node.

This component provides a powerful tool for visual navigation within the clothing section, allowing the user to discover combinations based on deep similarity relationships and obtaining more information about each of the different items.

E. Experimental Configuration

To carry out the implementation of the proposed system, different machine learning and natural language processing frameworks and tools were used. The main development environment was based on Python 3.10 due to the connection of this version with Rasa, using libraries such as TensorFlow, Keras, scikit-learn, Hugging Face Transformers, NetworkX, Plotly, Rasa and Streamlit.

Regarding image processing, a garment classification model was trained using the VGG16 architecture with pre-trained weights in ImageNet. The model was trained by freezing the convolutional base and adding custom dense layers for the classification of 11 garment categories.For style detection, a second model based on ResNet50 was used, also pre-trained and adapted with a final dense layer to classify 5 styles (casual, formal, sportive, elegant, urban).

The recommendation system was implemented using the *Surprise* library, using the *Singular Value Decomposition* (SVD) technique. Synthetic ratings were generated between users and garments, and cross-validation (*cross-validation*) was applied with 5 partitions to evaluate overall performance.

The natural language processing component was managed using the Rasa platform, combined with multilingual BERT models for intent detection. Text generation was carried out using the GPT-2 model through Hugging Face's Transformers library.

Metrics used to evaluate the system include:

- **Accuracy** of the garment classification model.

- **Precision** in the style prediction.
- **Top-K similarity** for visual recommendations.
- **RMSE** in the cross-validation of the collaborative system.
- **Conversational interaction and comprehension** in user tests.

All models and tests were carried out in a local development environment, using the Rasa platform as the core of the conversational system.

IV. Result Analysis

During the experimental phase, different tests were carried out to evaluate the performance of each module of the proposed system, ranging from garment classification to personalized recommendation based on image and text.

First, with the garment classification model implemented with VGG16, an accuracy of 99.08% was achieved, which indicates an excellent result in the task of identifying the type of garment from images. This confirms that, despite the small size of the dataset, the model was able to learn visual representations for all 11 categories.

On the other hand, the style classification model, based on a tight ResNet50 network, did not achieve such good results after 30 training periods. The accuracy obtained was 5.81%, which shows a high complexity in the task of classifying styles that is proposed as a line of improvement for future updates of the program.

Regarding the visual organization, the characteristics of the processed images were extracted to build a similarity graph, on which unsupervised clustering was applied, specifically KMeans, in which five main groupings were obtained. This graph made it possible to make recommendations based on visual similarity and facilitated the exploration of the garment space through search algorithms such as BFS. As for the collaborative recommendation system, it was implemented using the SVD algorithm on synthetic valuation data. The model obtained an average *Root Mean Square Error* (RMSE) of 1.4157 and a *Mean Absolute Error* (MAE) of 1.2170, which represents a good performance considering the ratings and the size of the set of users who submitted their opinions.

Finally, although no formal metrics of the conversational component or the generation of descriptions with GPT-2 are presented, both were evaluated qualitatively. During testing, the chatbot demonstrated the ability to correctly interpret intentions in most cases and generate consistent responses. The automatically generated outfit descriptions also proved to be understandable and useful to accompany the recommendations. Taken together, all these results show a functional and well-structured system with promising results, which has a small margin for improvement in terms of the classification of styles and the generalization of collaborative recommendations.

V. Conclusions and Future Work

The development of this project has demonstrated the feasibility of integrating different artificial intelligence techniques, such as natural language processing, computer vision, and recommendation systems, into a unified conversational assistance platform for the fashion world. The proposed architecture allows users to interact in a natural way, both through text and images, receiving personalized and visually coherent recommendations.

The system achieved high accuracy in garment classification using VGG16, as well as a graph-based visual recommendation structure that facilitates exploration between similar products. A collaborative recommendation engine was also incorporated, which, although based on synthetic assessments, showed consistent results. On a conversational level, the integration of Rasa allowed for fluid dialogues, while the generation of automatic descriptions added communicative value to the system. All together it showed a perfect combination of communication techniques and outstanding recommendations in a sector such as fashion.

However, it is true that there are some aspects in which certain improvements can be made. The model in charge of classifying the styles showed a poor performance, although favorable results have been concluded on the platform, which indicates the need to expand the dataset and work with examples. In addition, it would be interesting to evaluate the system with real users, collecting their feedback so that it allows you to measure their real opinions in a practical environment. As future work, it has been proposed to optimize the style module using more specific adjustment techniques, enriching the dataset with a greater number of images and adding real descriptions of greater diversity. Consideration has also been given to implementing the system on a web platform in production, allowing for open use and data collection to provide feedback on learning.

In short, this project is shown to be a solid basis for the creation of virtual assistants applied to the fashion sector, combining emerging technologies with an accessible, functional, and above all, visual interface.

References

[1] J. Iglesias, ”Human-like natural language interaction for service robots,” Ph.D. dissertation, Universitat Politècnica de Catalunya, 2015. [Online]. Available: https://www.tdx.cat/handle/10803/369847

[2] Papers With Code, “DeepFashion Dataset,” [Online]. Available: https://paperswithcode.com/dataset/deepfashion

[3] Teldat, “AI and BERT: Artificial Intelligence closer to human language,” 2020. [Online]. Available: https://www.teldat.com/es/blog/ai-inteligencia-artificial-bert/

[4] Papers With Code, “Multimodal Recommendation Task,” [Online]. Available: https://paperswithcode.com/task/multimodal-recommendation

[5] KeepCoding, "VGG16 and VGG19 architectures in Deep Learning," [Online]. Available: https://keepcoding.io/blog/arquitectura-vgg16-vgg19-deep-learning/

[6] Roboflow, "What is ResNet50?" [Online]. Available: https://blog.roboflow.com/what-is-resnet-50/

[7] Geeks for Geeks, "Singular Value Decomposition (SVD)," [Online]. Available: https://www.geeksforgeeks.org/singular-value-decomposition-svd/

[8] Broadcom, "Extending Python Agent with Rasa," [Online]. Available: https://techdocs.broadcom.com/es/es/ca-enterprise-software/it-operations-management/dx-apm-saas/SaaS/python-agent/Python-Agent-Extensions/Rasa.html

[9] Rasa Technologies, "Rasa Documentation," [Online]. Available: https://legacy-docs-oss.rasa.com/docs/rasa/

[10] Wikipedia, "Vanishing gradient problem," [Online]. Available: https://en.wikipedia.org/wiki/Vanishing_gradient_problem

[11] InteractiveChaos, "Singular Value Decomposition (SVD)," [Online]. Available: https://interactivechaos.com/es/wiki/descomposicion-en-valores-singulares-svd

[12] Datahack, "Recommendation engines with Python (Part 2)," [Online]. Available: https://www.datahack.es/motores-de-recomendacion-con-python-parte-2/

[13] Python Software Foundation, "Pickle — Python object serialization," Python 3.13 documentation, [Online]. Available: https://docs.python.org/es/3.13/library/pickle.html