Visual Recommendation and Visual Search for Fashion E-Commerce

Alessandro Abluton[1,2](✉)

[1] University of Turin, 10149 Turin, TO, Italy
alessandro.abluton@unito.it
[2] Inferendo s.r.l., 15121 Alessandria, AL, Italy

Abstract. Recommender systems are historically one of the most successfull and widely known applications of AI, personalized suggestions are nowadays a valuable commercial application of such systems. Many papers have been published in this field, but it is not yet solved; these models still lack state of the art multi-modal capabilities, such as conversational or visual suggestions. In this contribution we present a novel Visual Recommendation module for fashion e-commerces capable of recommending items based on a concept of visual similarity, and a Visual Search module where users can upload a picture of some clothing and search for the most similar ones in a given e-commerce. In conclusion we discuss about the accessibility of these recommender systems for small and medium enterprises, briefly describing our idea of Recommendations-as-a-Service.

Keywords: Recommender systems · Image similarity · Deep Learning

1 Introduction

Few applications of Artificial Intelligence have seen as much commercial success as Recommender Systems; user-tailored suggestions are nowadays present in almost every aspect of our interactions with e-commerces, streaming services, newscasts, social networks and are about to enter the Web3.0 and Metaverse era. Given the astonishingly good results that such systems provided over the years it is not surprising that research on this field is flourishing and interest by both academic and industrial players is growing rapidly.

The latest trends in Recommender Systems [1] concern mostly on pushing the frontiers in several open challenges regarding conversational systems, fairness in recommendations, evaluation methods and general domain-specific enhancements. In this last context a new clear trend is the development of Visual Recommenders based on modern neural models, able to take into account the visual features of an item to make suggestions; the need for new kinds of interactions mechanism between users and e-commerces is rising, due to both a new generation of customers and to the amazing enhancements in neural networks built to handle image data.

T. Skopal et al. (Eds.): SISAP 2022, LNCS 13590, pp. 299–304, 2022.
https://doi.org/10.1007/978-3-031-17849-8_25

Another interesting topic that has not been addressed by the research community is the accessibility of these kind of services, the focus is mostly on implementing complex and resource-hungry Deep Learning systems to further enhance the quality of the suggestions. No effort is put into allowing these kind of systems to also reach Small and Medium Entreprises (SMEs), who lack resources to implement such complex systems in terms of money and data availability.

2 Visual Recommendation and Visual Search

Even if traditional recommendation approaches have proven to be accurate, efficient and easy to implement, the need for new solutions for interactions between users and e-commerces is rising. New generation of consumers are accustomed to much more complex interactions with their devices, platforms that will implement them will be rewarded on the long run.

We built a system able to perform image similarity search on top of any kind of fashion e-commerce in order to recommend the most similar-looking products. We also built an object detection module in order to add visual search capabilities to the system.

2.1 Image Similarity

The feature representation and similarity measurement, which have been thoroughly explored by multimedia researchers for decades, are key components that determine how well a content-based image retrieval system performs, this continues to be one of the most complex issues in current content-based image retrieval (CBIR) research despite a range of solutions being suggested.

Following the ideas presented by Wan et al. [3] we chose to represent item's images with embeddings in a latent space, obtained by extracting the activations of the last layers of a Convolutional Neural Network. The idea is simple, first we train a CNN on a traditional classification task, so that the network can learn to position similar items (e.g. belonging to the same class) nearby one to another in the latent space, then *K-NN* based search can be used to find the most similar images to a given one.

The currently implemented search is based on the *cosine distance* metric, but we plan to take into account all the tecniques available in the literature.

2.2 EfficientNet Models

The decision on which model to use led us to search for the most modern convolutional architectures available in literature, we settled on the EfficientNet [4] family of networks and we are currently still experimenting on which particular implementation to use due to the availability of 7 different EffNets (B0 to B7) with increasing levels of complexity and image resolutions they operate on. This choice depends on the typical trade-off between representation capabilities and time required to effectively train the model on a real-case dataset and to make predictions in a production environment.

We trained two different version of the same EfficientNet model: for the tasks of visual recommendation we used a dataset made of whole products images that exploits also the context around a product (dresses of different lenghts are easily distinguished when the whole figure is visible) and for visual search a dataset made of cropped products images has been used, to reduce the model context dependency. Figure 1 shows a graphical explanation.

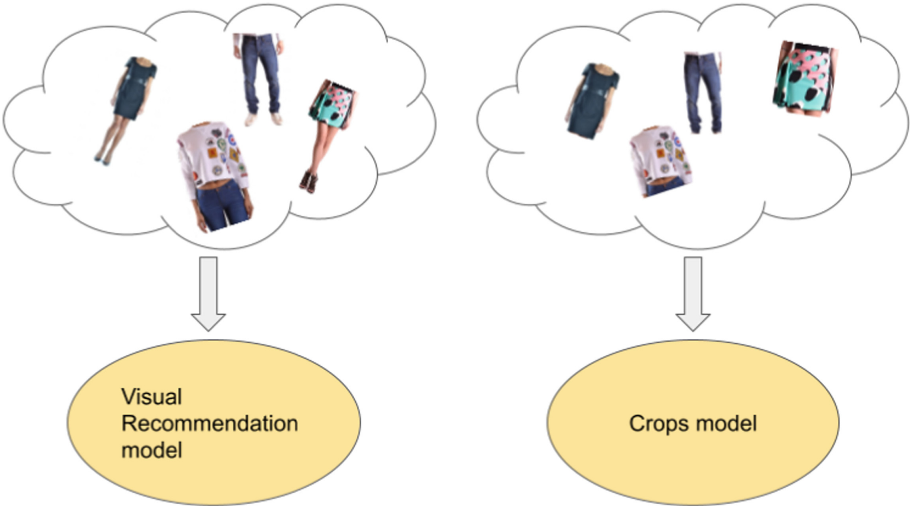

Fig. 1. Two different models are used

2.3 Visual Recommendation

In the context of fashion e-commerces *Visual Recommendation* means to be able to suggest to users the most **visually similar** items in respect to the one they are looking at. Let us consider the example situation of a user viewing the page of a Dress, in the context of visual recommendation the first image of that product (usually the most representative one) is given in input to the embedding model and an exact *K-NN* search will be performed between this and all the other images of products belonging to the same class, so for this example only dresses embeddings will be queried, Fig. 2 shows an actual example on a small e-commerce dataset.

2.4 Visual Search

Plain visual recommendations are not enough to achieve the goal of innovative interactions methods between customers and e-ecommerces, being just

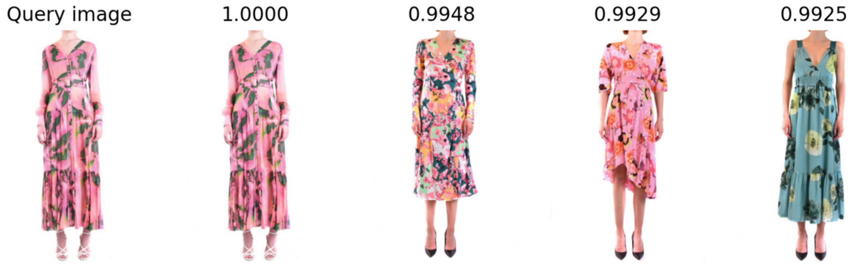

Fig. 2. An example of visual recommendations from a dress item, similarity score is shown on top of each image

another way of suggesting products to customers. We decided to implement a *Visual Search* module inspired by the functionalities of Google Lens (https://lens.google/), a service that allows users to upload a picture of any kind and to search the web for items inside of it. Figure 3 shows a complete high-level representation of this process. We implemented a plugin available for

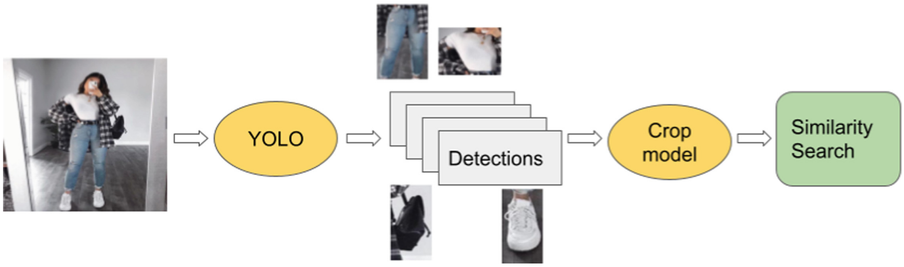

Fig. 3. High-level schema of a complete visual search process

the major e-commerce platforms in the market (Shopify, Magento, Woocommerce) that can enable this feature in any kind of fashion e-commerce. Figure 4 shows an example of the beta implementation currently available in production: users can upload a picture, items inside will be evidenced by the grey dots, when they click on one of the items the actual crop will be embedded by using the *Crop model* and the result sent to the visual recommendation module, in order to find the most similar ones in that e-commerce.

Objects inside of an image are recognized by means of the Yolo (You only look once) model [5], a well known object-detection model trained on the Open-Images dataset provided by Google [6] from which we downloaded all the images concerning the clothing categories and their relative bounding boxes.

Handling Any Fashion E-Commerce. The biggest challenge we are facing is generality with respect to all the possible categories that can be found in an e-commerce, as stated in Sect. 2.1 the convolutional network is trained on a fixed set of classes, but fashion e-commerces tend to have very different categorization of clothing with a considerable rate of errors both in terms of misplaced product classes and grammatical errors in the actual names of said categories.

To handle these problems we decided to settle on a fixed set of classes defined by us, we map each product of any e-commerce that uses our plugin on that set of classes by exploiting the classification capabilities of our CNN model. The process of training is as follows:

1. Start from a pretrained EfficientNet (on imagenet dataset).
2. Build a training dataset with our fixed set of classes, by means of manual web crawling.
3. Fine-tune the EffNet on that dataset for a classification task.

We then obtain a general model able to classify any kind of clothing picture into our set of classes, therefore able to produce embeddings with enough representative power to perform visual recommendation on any fashion e-commerce, without the need of additional fine-tuning steps for each new client.

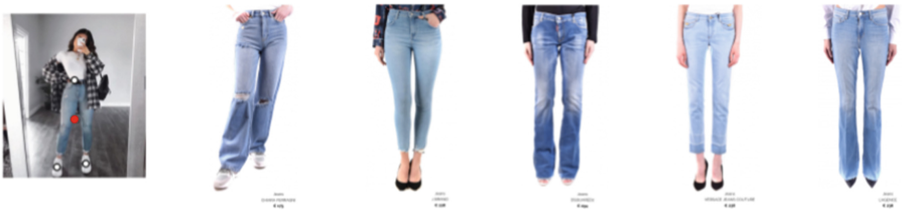

Fig. 4. The visual search module, users can upload a picture and select items to search

3 Conclusions and Future Work

We presented a novel method for implementing visual search and recommendation on any kind of fashion e-commerce without the need of fine tuning steps and through the idea of Recommendations-as-a-Service. We built a cloud infrastructure that serves as backend for the plugins we developed, enabling virtually any kind of e-commerce on major platforms to adopt these new interactions by just installing a plugin, thus enabling SMEs with low economic budget and development resources to access these new technologies.

As future work, we plan to further develop the RaaS infrastructure in a scalable and modular way, in order to address the problem of SMEs not having enough resources to actually use these new technologies; we believe there is a niche of market yet to be filled in this regard.

Further and extensive experimentation is also still needed to enhance the quality of the visual recommenders we presented, by building better training datasets and via the implementation of some sort of continuos learning mechanism, taking into account users feedbacks on our suggestions. Finally, we would like to address some of the latest trends in recommender systems, with a particular focus on conversational and multi-modal algorithms.

References

1. Jannach, D., Pu, P., Ricci, F., Zanker, M.: Recommender systems: trends and frontiers. AI Mag. **43**, 145–50 (2022). https://doi.org/10.1002/aaai.12050
2. Koren, Y., Bell, R.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook, 2nd edn., pp. 77–118 (2015)
3. Wan, J., et al: Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 157–166 (2014)
4. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp. 6105–6114 (2019)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection (2015). CoRR. http://arxiv.org/abs/1506.02640
6. Kuznetsova, A., et al.: The open images dataset V4. Int. J. Comput. Vision **128**(7), 1956–1981 (2020). https://doi.org/10.1007/s11263-020-01316-z