“이거 어디서 사?” – Mask R-CNN 기반 객체 분할을 활용한 패션 아이템 검색 시스템

정경희 [1], 최하늘 [2], Sammy Y. X. B.[1], 김현성 [3], N. D. Toan[1], 추현승 [3]
[1] 성균관대학교 수퍼인텔리전스학과
[2] 성균관대학교 소프트웨어학과
[3] 성균관대학교 전기컴퓨터공학부
datakira@g.skku.edu, choo@skku.edu

“Where can I buy this?” - Fashion Item Searcher using Instance Segmentation with Mask R-CNN

Kyunghee Jung[1], Ha nl Choi[2], Sammy Y. X. B.[1], Hyunsung Kim[2], N. D. Toan[1], Hyunseung Choo[3]

[1]Dept. of Superintelligence, Sungkyunkwan University
[2]College of Software, Sungkyunkwan University
[3]Dept. of Electrical and Computer Engineering, Sungkyunkwan University

Abstract

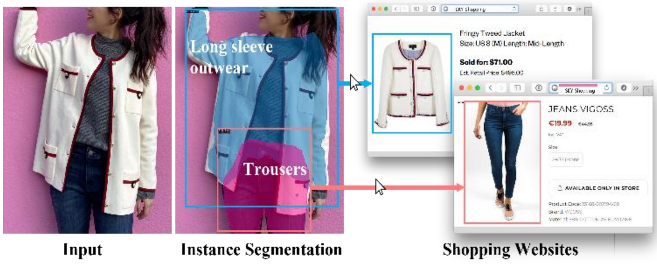
Mobile phones have become an essential item nowadays since it provides access to online platform and service fast and easy. Coming to these platforms such as Social Network Service (SNS) for shopping have been a go-to option for many people. However, searching for a specific fashion item in the picture is challenging, where users need to try multiple searches by combining appropriate search keywords. To tackle this problem, we propose a system that could provide immediate access to websites related to fashion items. In the framework, we also propose a deep learning model for an automatic analysis of image contexts using instance segmentation. We use transfer learning by utilizing Deep fashion 2 to maximize our model accuracy. After segmenting all the fashion item objects in the image, the related search information is retrieved when the object is clicked. Furthermore, we successfully deploy our system so that it could be assessable using any web browser. We prove that deep learning could be a promising tool not only for scientific purpose but also applicable to commercial shopping.

1. Introduction

Shopping online has become a habit for many people, especially with the development of technology and internet. However, it is not always the case that we could find the stores of a wanted items, especially when we see the item somewhere on the internet. In order to obtain optimal search information related to a specific fashion item in the picture, it is necessary for the user to directly try multiple searches by combining appropriate search keywords. Fashion Items Searcher analyses the context of photos, divides them into object units, and extracts keywords and cropped images for each object. We adopt the popular instance segmentation method to detect an item within an image and then use these output masks to redirect users to the shopping website.

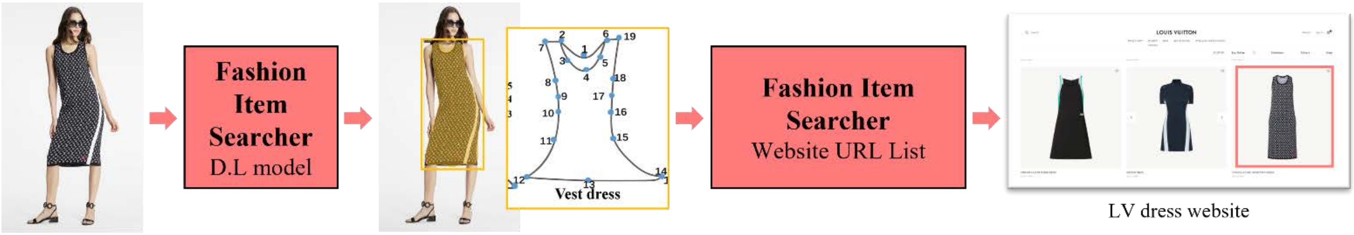
2. Related Work

Image segmentation is the task of partitioning an image into different masks, representing the corresponding image objects. The wide range application of this technique makes it become a hot topic in deep learning field, including medical imaging [1], self-driving vehicles [2] or video surveillance [3]. Instance segmentation [4][5] is an extension of image segmentation, where different objects of similar type appeared in the images, but isolated segments are needed.



[Fig 1] Illustration of how our system works

Many models have proposed to tackle this problem [6][7], with one of the most successful core structures is Mask R-CNN. As the name suggests, Mask R-CNN [8] use Convolutional Neural Network to learn the features of the images, enable the recognition of objects. Built upon that, R-



[Fig 2] Illustration of how Fashion Items Searcher works

CNN (stands for Region-Based CNN) was introduced to evaluate the Regions of Interest (ROI) in the images based on the bounding box. This improves the model as specific regions are taken into account in inferencing steps. Finally, Mask R-CNN bring about the mask for each of the object, providing a more profound understand of the network.

Image classification in fashion platform using deep learning has been challenging task in computer vision. Fashion-MNSIT dataset [9] is one of the most famous fashion datasets containing a training set of 60,000 examples with 10 categories. Existing deep learning models for classifying fashion items are trained as transfer training with pre-trained model. State-of-art deep learning methods were used to tackle this problem. [10] trained deep learning model using fine-tuning DARTS with fixed operations resulting top-1 accuracy 96.91% for classifying clothes on Fashion-MNIST dataset.

3. Methodology

Our system uses instance segmentation as a core method for detecting fashion items, where object recognition and segmentation are implemented to achieve this. We adopt the well-known Mask R-CNN deep learning models, which has demonstrated an outstanding performance in instance segmentation, to learn the feature representations of the images. The process is as follows:

1. The model is pretrained with COCO dataset [11] using Detectron2, gaining an initial knowledge about images.
2. Using transfer learning, the model is trained again using Deep Fashion 2 dataset, learning a better understanding of fashion items in images.
3. The model is deployed using Amazon Web Service and Flask, so that can be accessible in web browser.
4. Finally, Beautiful Soup with Google API is used with our model to search for the website that sell the items.

3.1 Dataset

Deep Fashion2 is a fashion dataset that can be used publicly with rich annotations. It contains 13 cloth classes and about 300 landmark information, and is divided into about 39,100 learning images, 34,000 validation images, and about 67,000 test images. Each image is labeled with size, degree of occlusion, scale, perspective, category, style, bounding box, landmark, and mask by pixel. [10] It contains the information necessary to implement our system, so it is regarded to be suitable for use in training the Mask R-CNN model.

3.2 Model

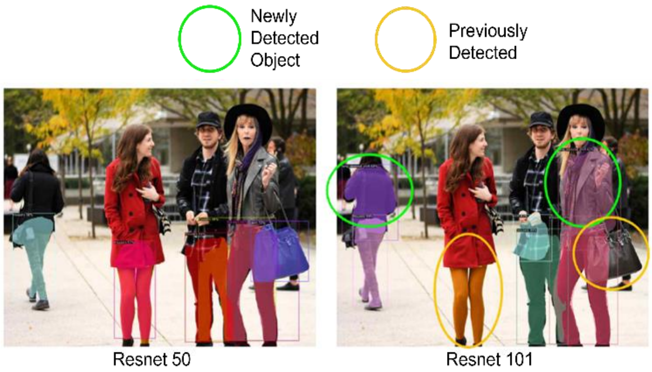
To detect fashion elements in an image, we implement instance segmentation technique, which classifies overlapping parts in pixels and detects all objects in the image. It takes advantages for distinguishing instances of the same classified category. It is combined with object detection and semantic segmentation. For example, if there are five dresses in an image, each five dresses is recognized as a different instance even though they are all classified into the same class.

Detectron2 is a learning and inference platform for object detection and semantic segmentation developed by Facebook Artificial Intelligence Research (FAIR). Since it is based on the Mask R-CNN model, it provides functions mainly used such as bounding box, segmentation, and key point.

When an image is fed into model, object detection and instance segmentation are performed parallelly. They are classified and keywords are extracted. Furthermore, the information such as coordinates and silhouettes for each object are displayed with the keywords.

4. Result

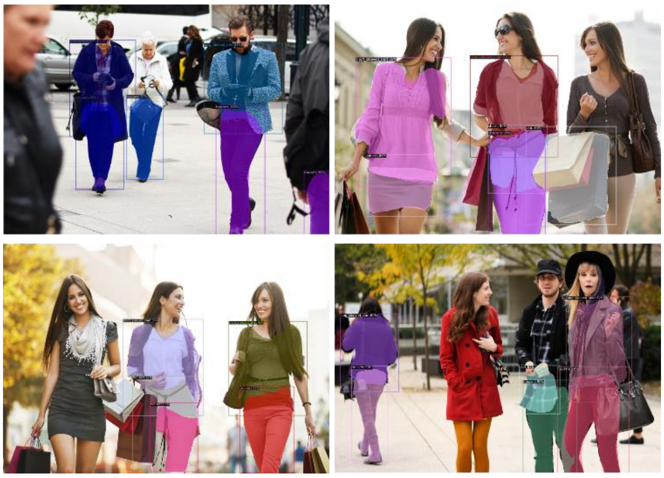
In order to achieve the best results, we carry out experiments on different models as a backbone of the instance segmentation. Fig 2 illustrates an example of our trial, where we use the same image as an input for 2 different classifiers.



[Fig 3] Instance segmentation output of using ResNet50 and ResNet101 as backbone

The image on the left-hand side if the output of ResNet50, and the right-hand side is ResNet101. Compared to ResNet50, ResNet101 have detected more objects, for example the long sleeve shirt. Similar experiments have been carried out, and finally ResNet101 have achieved the best performance, register it as a core model for Mask R-CNN.

Demonstrated in Fig 3, the model outputs the segmented instances of fashion items. After the model analyses the context of photos, multiple keywords and coordinates of the instances are extracted. Then using the coordinates, we crop the instances and save them with labels. The labels for each item include: short sleeve top, sleeve top, short sleeve outwear, long sleeve outwear, vest, sling, long sleeve dress, vest dress, sling dress, shorts, trousers, skirts, short sleeve dress.



[Fig 4] Sample outputs of Fashion Item Searcher deep learning model

After the segmentation mask is obtained, these outputs are fed into search engines such as Google API and returns the most similar link among the results, as shown in Fig 1.

5. Conclusion

Fashion Item Searcher automatically analyses an image context extracting keywords and segmented object images through deep learning. After segmenting all the fashion item objects in the image, the related search information is retrieved at once when the object is clicked.

Our framework is also able to be combined with various fields such as shopping, social network services, and e-commerce markets. Based on the history data containing clicked and searched fashion items, personalized advertisements can be recommended. Furthermore, a partnership service and specific shopping websites related to the fashion items can be preferentially displayed to users.

Acknowledgement

This research was supported by National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2020R1A2C2008447), under the ICT Creative Consilience program(IITP-2022-2020-0-01821), and Grand Information Technology Research Center support program(IITP-2022-2015-0-00742) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

Reference

[1] Lai, Matthew. "Deep learning for medical image segmentation." arXiv preprint arXiv:1505.02000 (2015).

[2] Treml, Michael, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr et al. "Speeding up semantic segmentation for autonomous driving." (2016).

[3] Sun, Hongzan, Tao Feng, and Tieniu Tan. "Spatio-temporal segmentation for video surveillance." In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 1, pp. 843-846. IEEE, 2000.

[4] De Brabandere, Bert, Davy Neven, and Luc Van Gool. "Semantic instance segmentation with a discriminative loss function." arXiv preprint arXiv:1708.02551 (2017)

[5] De Brabandere, Bert, Davy Neven, and Luc Van Gool. "Semantic instance segmentation for autonomous driving." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 7-9. 2017.

[6] Wei, Yixuan, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. "Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation." arXiv preprint arXiv:2205.14141 (2022).

[7] Mayer, Zoe, J. Kahn, Y. Hou, and R. Volk. "AI-based thermal bridge detection of building rooftops on district scale using aerial images." In Proceedings of the EG-ICE 2021 Workshop on Intelligent Computing in Engineering proceedings, Berlin, Germany, vol. 30. 2021.

[8] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

[9] Xiao, Han, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms." arXiv preprint arXiv:1708.07747 (2017).

[10] Tanveer, Muhammad Suhaib, Muhammad Umar Karim Khan, and Chong-Min Kyung. "Fine-tuning darts for image classification." In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4789-4796. IEEE, 2021.

[11] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In *European conference on computer vision*, pp. 740-755. Springer, Cham, 2014.