Object detection and instance segmentation of fashion images

Sandra Treneska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius, Skopje, Macedonia
sandra.treneska@students.finki.ukim.mk
Skopje, 2020

Sonja Gievska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius, Skopje, Macedonia
sonja.gievska@finki.ukim.mk
Skopje, 2020

***Abstract*—Over the past few years, fashion brands have been rapidly implementing computer vision into the fashion industry. Our research objective was to analyse a number of methods suitable for object detection and segmentation of apparel in fashion images. Two types of models are proposed. The first, simpler, is a convolutional neural network used for object detection of clothing items on the Fashion-MNIST dataset and the second, more complex Mask R-CNN model is used for object detection and instance segmentation on the iMaterialist dataset. The performance of the first proposed model reached 93% accuracy. Furthermore, the results from the Mask R-CNN model are visualized.**

***Index Terms*—object detection, instance segmentation, semantic segmentation, computer vision, fashion images**

I. Introduction

More than 25% of the entire revenue in E-Commerce is attributed to apparels and accessories. A major problem they face is categorizing these apparels from just the images. This poses an interesting computer vision problem.

Analyses of fashion images are popular research topics in recent years because of their huge potential in the industry. Detection of clothing items from a single image can have huge commercial and cultural impact on society. Many researches in this field have recently progressed from recognition-based clothing retrieval tasks to understanding-based tasks. That means that the models can not only recognize the attributes of fashion images but can also understand the meaning of the combination of those attributes.

Object detection and segmentation have a wide spectrum of computer vision applications for fashion, including online shopping, personalized recommendation and virtual try-on. Many fashion brands are already using machine learning techniques to predict and design what will be the next fashion trend [1] or for visual search [2].

However, real-world application remains a challenge, because of deformations, occlusions and discrepancies between consumer and commercial clothing images. Also, problems may occur due to wide variations in appearance, style, brand and layering of clothing items. At the same time, very subtle differences can exist that cannot be easily distinguished, for example images of the same product can often look different under different conditions. Therefore, these tasks are being extensively studied in computer vision community by many research groups in both academia and industry.

Instance segmentation is challenging because it requires the correct detection of all objects in an image while also precisely segmenting each instance. It combines traditional object detection and semantic segmentation. The goal of object detection is to classify individual objects and localize each using a bounding box. The goal of semantic segmentation is to classify each pixel into a fixed set of categories.

II. Related work

The DeepFashion2 paper [3] is a benchmark for detection, pose estimation, segmentation and re-identification of clothing images. They try to fill the gap of a previous paper, Deep Fashion and address its issues including single clothing item per image, sparse landmarks, and no per-pixel masks. They also propose a baseline, Match R-CNN, which builds upon Mask R-CNN and solves the issues in an end-to-end manner.

ModaNet [4] provides a dataset of street images fully annotated with masks (polygons) of a single person. ModaNet aims to provide help for evaluating the progress of the latest computer vision techniques that rely on large data for fashion understanding. The rich annotation of the dataset allows to measure the performance of algorithms for object detection, semantic segmentation and polygon prediction of images in detail.

FashionAI [5] presents a hierarchical dataset for fashion understanding. They realize that fine-grained attribute recognition is critical, but it was missing form existing datasets. FashionAI addressed this by building a well-structured hierarchical knowledge and precise annotations of fashion apparel.

III. Models

A. Datasets

Fashion-MNIST [6] is a dataset of article images consisting of a training set of 60,000 examples and a test set of 10,000

examples. Each example is a 28x28 gray-scale image, associated with a label from 10 classes. The label descriptions are the following: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot.

iMaterialist [7] dataset is provided by FGVC6 workshop at the Conference on Computer Vision and Pattern Recognition (CVPR) 2019. It contains a total of 50,000 clothing images from daily-life, celebrity events, and online shopping. Their taxonomy contains 46 apparel objects (27 main apparel items and 19 apparel parts), and 92 related fine-grained attributes.

B. CNN model

In this part, a CNN model is trained on the Fashion-MNIST dataset for the purpose of classifying fashion items in images.

For the preprocessing of the images, the pixels are normalized and then every image is reshaped as a numpy array of pixels. Resizing the images is not necessary since all the images are already the same size. Next, the data is split into train, validation and test sets and we have 48.000, 10.000 and 12.000 images in each set respectively.

Three CNN models of different complexity are created. The first one has one convolutional layer with 32 filters which results in 173,738 trainable parameters. The second model has two convolutional layers with 32 and 64 filters, resulting in 515,146 trainable parameters. Finally, the third model has three convolutional layers with 32, 64 and 128 filters resulting in 1,421,194 parameters. All the models have dropout layers which help the models to not overfit. At the end, all models have two dense layers, one with ReLU and one with softmax activation function. For training, sparse categorical cross entropy is used and an Adam optimizer.

Kaggle notebooks are used for the training, so the computations are done faster, on a GPU.

C. Mask R-CNN model

In this part, a Mask R-CNN model [8] with COCO pre-trained weights [9] is trained on the iMaterialist dataset for the purpose of object detection and instance segmentation of fashion images.

Mask R-CNN is a recent advanced framework developed by FAIR (Facebook AI Research) for object instance segmentation. It detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. It stands for Mask Regional Convolutional Neural Network and it extends Faster R-CNN. Additionally, Mask R-CNN is easy to generalize to any task and can also be used for keypoint detection.

The model can be roughly divided into 2 parts — a region proposal network (RPN) and binary mask classifier. The first step sets bounding boxes that could possibly contain an object of relevance, this is called ROI (Region of Interest) Align. These boxes are then refined using a regression model. In the second step instance segmentation is applied to each box. The instance segmentation model is trained like a binary classifier, meaning 1 represents the presence of an object and 0 represents the background. The architecture of Mask R-CNN is shown on Fig 1.

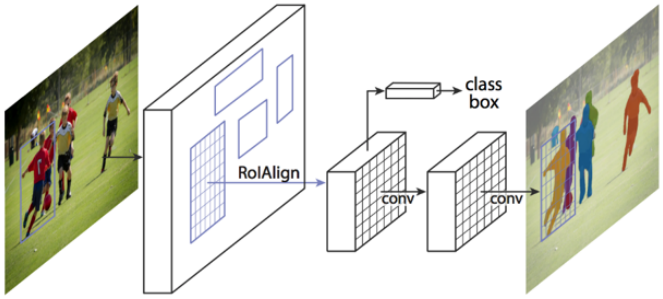

Fig. 1. Diagram of Mask R-CNN

For the configuration of the model, the steps per epoch and validation steps are lowered so the training can finish in less than a day. Also the image size is set to 512x512 and all images are resized to those dimensions. All the other hyperparameters were left as default.

The images have masks containing the pixels where the fashion items are. One image can have multiple masks, meaning that this a multi-label problem. Some of the masks are visualized, shown on Fig 2.

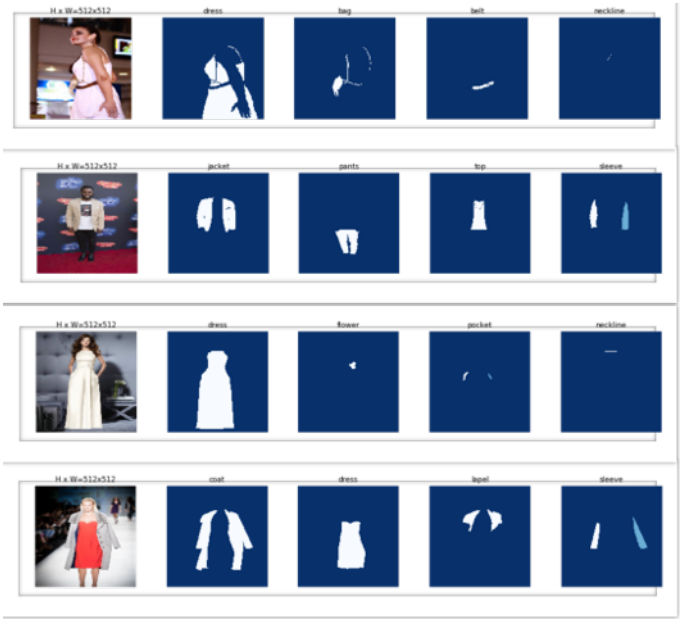

Fig. 2. Masks

Again, Kaggle notebooks were used for the purpose of training the model faster on a GPU. The training was done on 36,156 images that contain 264,949 segments (masks) in total, while the validation set had 9,039 images and 66,264 segments.

IV. Results

Accuracy and loss were measured for all the three CNN models. The results can be seen below on Fig 3 and Fig 4 for 10 and 50 epochs respectively.

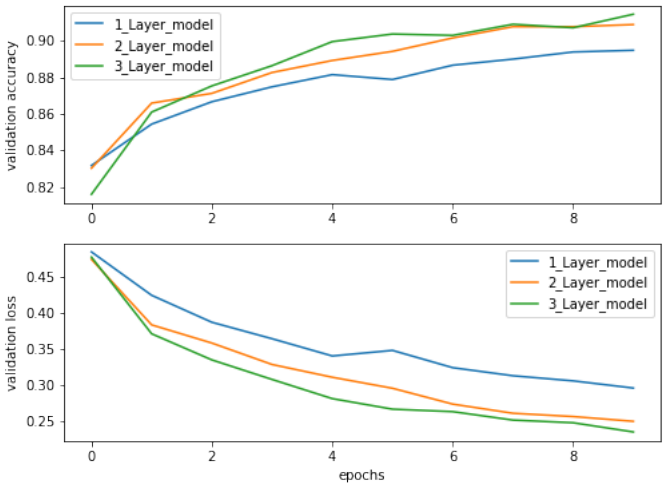

Fig. 3. 10 Epochs accuracy/loss graph

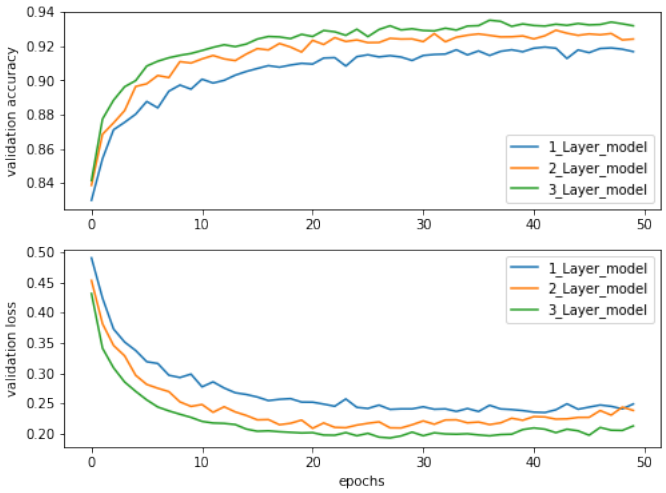

Fig. 4. 50 Epochs accuracy/loss graph

After 10 epochs we can say that there isn't much difference in the performance of the second and third model. But, in the long run we can see that the third model outperforms the others and after 50 epochs it has 93% accuracy. That was expected, since the third model has over a million parameters and naturally it needs more time to train.

The training accuracies for the first, second and third model respectively are 0.916, 0.928, 0.934. The F1-scores, micro and macro, are the same as the accuracies.

Since iMaterialist is a competition, the real classes of the test images weren't provided, so evaluation metrics couldn't be used for the Mask R-CNN model. But, it was still possible to visualize the predictions that the model made for the test images. Below, on Fig 5 are the testing images visualized.

For each image the model predicts the class, a bounding box, a mask and a confidence factor of every predicted fashion item. The model performed the best on images where the person is facing straight to the camera, and it was less accurate on sideways pictures or images where the clothes are captured from a different angle.

V. Conclusion and Future Work

The fashion industry has lately attracted a lot of attention with its huge economic potential and practical value. There are many researches and competitions that analyze how computer vision can be integrated in the fashion industry.

In this paper the focus was on object detection and instance segmentation of fashion items in images. By using convolutional neural networks and Mask R-CNN for these tasks we were able to produce meaningful results.

Future work could further improve the model's performance. Both models can be improved with more training time and computational resources as well as parameters tuning and data augmentation techniques. Additionally, classifying fine-grained attributes for the fashion items in iMaterialist dataset could be implemented.

References

[1] Stitch Fix https://algorithms-tour.stitchfix.com/
[2] Pinterest visual serch https://newsroom.pinterest.com/en/post/introducing-the-next-wave-of-visual-search-and-shopping
[3] Ge, Yuying, et al. "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[4] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In ACM Multimedia, 2018.
[5] Zou, Xingxing, et al. "FashionAI: A Hierarchical Dataset for Fashion Understanding." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.
[6] Fashion-MNIST dataset https://github.com/zalandoresearch/fashion-mnist
[7] iMaterialist dataset https://sites.google.com/view/fgvc6/competitions/imat-product-2019
[8] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
[9] COCO http://cocodataset.org/

Fig. 5. Predictions from the Mask R-CNN model